2017

Leveraging Structural Flexibility to Predict Protein Function

Ziyi Guo
Lehigh University

Leveraging Structural Flexibility to Predict Protein Function

by

Ziyi Guo

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

Computer Science

Lehigh University

August 2017

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Date

Committee Members:

Brian Chen, Committee Chair

Ting Wang

Xiaolei Huang

Katya Scheinberg

Acknowledgements

I would like to show my deepest gratitude to my Ph.D. advisor, Prof. Brian Chen, for his guidance in past five years. I am especially grateful for his patience in training me to be an independent researcher and a qualified collaborator, the freedom he gave in Informatics Lab and caring beyond academia. Without his efforts, this work is definitely impossible. I would like to thank my committee members, Prof. Ting Wang, Prof. Xiaolei Huang and Prof. Katya Scheinberg, for their precious suggestions on my Ph.D. proposal, general exam and dissertation. I would also like to show my thanks to Prof. Soutir Bandyopadhyay and Prof. Katya Scheinberg whom I have worked with at Lehigh. Prof. Bandyopadhyay provided insightful comments on statistical modelling for protein structure comparison and Prof. Scheinberg's expertise in mathematical optimization greatly helped our work on protein electrostatic analysis.

I am tremendously thankful to my friends at Lehigh, Yuhai Hu, Mengtao Sun, Wenjia Ruan and Yujie Liu. Studying as a Ph.D. student in a foreign country is never easy, full of challenges, failure and depression. Life in the past five years could be more difficult without countless help and sharing from these friends and I have always enjoyed to talk to them, wherever it was, Saxbys Coffee at campus or Izakaya restaurants in NYC.

Lastly, thanks to my parents, Shaojian Guo and Chunxiao Li, should be never-ending for their unconditional support and love.

The work presented in this thesis is supported in part by National Science Foundation Grant NSF-1320137. The experiments were made possible with Corona and

Sol servers, Lehigh University.

Contents

List of Tables

List of Figures

Abstract

Proteins are essentially versatile and flexible molecules and understanding protein function plays a fundamental role in understanding biological systems. Protein structure comparisons are widely used for revealing protein function. However, with rigidity or partial rigidity assumption, most existing comparison methods do not consider conformational flexibility in protein structures. To address this issue, this thesis seeks to develop algorithms for flexible structure comparisons to predict one specific aspect of protein function, binding specificity. Given conformational samples as flexibility representation, we focus on two predictive problems related to specificity: *aggregate prediction* and *individual prediction*.

For aggregate prediction, we have designed FAVA (Flexible Aggregate Volumetric Analysis). FAVA is the first conformationally general method to compare proteins with identical folds but different specificities. FAVA is able to correctly categorize members of protein superfamilies and to identify influential amino acids that cause different specificities. A second method PEAP (Point-based Ensemble for Aggregate Prediction) employs ensemble clustering techniques from many base clustering to predict binding specificity. This method incorporates structural motions of functional substructures and is capable of mitigating prediction errors.

For individual prediction, the first method is an atomic point representation for representing flexibilities in the binding cavity. This representation is able to predict binding specificity on each protein conformation with high accuracy, and it is the first to analyze maps of binding cavity conformations that describe proteins with different specificities. Our second method introduces a volumetric lattice

representation. This representation localizes solvent-accessible shape of the binding cavity by computing cavity volume in each user-defined space. It proves to be more informative than point-based representations. Last but not least, we discuss a structure-independent representation. This representation builds a lattice model on protein electrostatic isopotentials. This is the first known method to predict binding specificity explicitly from the perspective of electrostatic fields.

The methods presented in this thesis incorporate the variety of protein conformations into the analysis of protein ligand binding, and provide more views on flexible structure comparisons and structure-based function annotation of molecular design.

Chapter 1

Introduction

1.1 Motivation

Protein functions refer to all types of biochemical activities that a protein play in biological systems, and they are essential to living organisms. For example, antibodies bind to specific particles to protect human body [4]. Messenger proteins transmit signals to coordinate biological activities between different cells, tissues and organs [5]. With extensive knowledge of protein function, many practical goals, such as boosting identification of drug targets, reducing side effects in protein engineering and designing synthesis of new types of bio-materials, can be achieved. In addition, understanding protein function is fundamental to expanding biological research. Studies on protein-protein interactions, protein-DNA/RNA interactions and protein network construction can be accelerated in a broader way with understanding function of individual proteins.

A variety of genome sequencing and structural genomics projects provide us an increasing amount of high throughout protein sequences and structures. Unfortunately, function of many proteins is not available yet. About 40% of protein structures in the NCBI database, for example, are not assigned with functional information [6]. To close the gap between available protein sequences/structures and unknown protein function, biologists carry out biochemical experiments to deter-

mine the function of each protein. However, these experimental efforts are costly and time-consuming and cannot automate the elucidating of protein function because they highly rely on the insights of skilled biologists and involve in a large-scale experimentation. Hence, developing computational methods for automatic protein function annotation is a challenging problem in modern bioinformatics.

1.2 What Exactly is Protein Function?

Proteins are extremely versatile and the concept of protein function is highly context-sensitive and is not always well-defined. Therefore, to refine the protein function definition, this section includes a discussion of protein function from various perspectives and many efforts that standardize protein function description.

Protein function can be understood from different levels: from specific molecular biochemical reactions up to actions of the organism as a system. Here, we take a categorization of the types of protein function that is suggested by Bork et al. [7].

- **Molecular Function**: The biochemical function performed by a single protein, such as ligand binding, biochemical reaction catalysis and conformational changes.

- **Cellular Function**: Multiple proteins work jointly to perform more complex physiological functions, such as metabolic pathways, signal transduction and cellular localization , to keep a specific component of the organism in good condition.

- **Phenotypic Function**: With integration of biological stimuli from the environment, many physiological subsystems come together to determine phenotypic properties of the organism.

From the above discussion, protein function appears to be a very subjective concept. To standardize protein function definition, many protein classification systems have been proposed. One early work Enzyme Classification (EC) [8] focused on

the classification of enzymes which are macromolecules for catalysing biochemical reactions. The EC was proposed by the International Union of Biochemistry and Molecular Biology in 1992, and it provides a hierarchical classification of enzymes based on the chemical reactions they catalyze. The EC classification is a sequence of four numbers separated by periods that gives a progressively finer definition of a specific enzyme family: from the class of the reaction, the substrate, the type of chemical bonds to other binding specificities. Nevertheless, EC has a limited scope where only enzyme proteins get defined and classified. Many other similar classification, subsequent to EC, were proposed for a wider scope of function definition. However, they focused on specific organisms, such as EcoCyc [9] for E. coli genes and SubtiList [10] for B. subtilis genes. One general function description system is Gene Ontology (GO) [11]. Go provides an ontology of defined terms representing gene product properties that covers three parts: cellular component, molecular function and biological process. GO uses controlled vocabularies and is machine-readable, and has been recognized as the most commonly used system for functional annotation.

1.3 The Problem of Protein Function Prediction

From the above discussion, it is obvious that protein function can be understood from multiple perspectives, and this variation will generate a wide variety of biological data. Depending on the type of biological data, computational methods for protein function prediction are greatly diversified. Thus, this section discusses different categories of function prediction methods. After that, we introduce the specific problem studied in this thesis with highlight on how it is different from existing works.

Protein function prediction assigns specific information that indicates biological or biochemical roles to proteins. Protein function prediction methods take the following interface:

- **Input**: A given protein **P** and **K**, any extra biological knowledge about **P**.

- **Output**: Specific information at the level of a defined biological function.

Function prediction methods can be categorized depending on the data type of $\mathbf{K}$, such as genomic sequences [12, 13, 14, 15], phylogenetic profiles [16, 17, 18], gene expression [19, 20], protein interaction networks [21, 22] or even text mining from literature [23, 24]. Unfortunately, in many cases, extra information about the input protein is not always available. Computational methods that only take $\mathbf{P}$ as input generally involve two main types: the sequence-based approaches and the structure-based approaches. The sequence-based methods [25, 26, 27, 28, 29] compare one protein with unknown function to another with known function in the database, such as GenBank [30], in search of sequence similarity that is sufficient to indicate similar function. It was found that homologous proteins with more than 40% sequence identity tend to have identical function [31]. However, there always exist exceptions when the similarity is below 40% and only sequence is not robust enough for function prediction. It is well-known that protein structures are more evolutionarily conserved than protein sequences [32, 33, 34], and thus protein structures could be better predictive markers to be related to protein function, leading to structure-based methods.

The Protein Data Bank [35] is by far the most comprehensive repository of experimentally determined protein structures. As of April 2017, Protein Data Bank contains about 129,000 structures that are assigned by various experimentations, such as X-ray crystallography, NMR spectroscopy and electron microscopy. The $3D$ structure of each protein comes in the form of PDB file where the coordinate of each atom is recorded. Structure-based methods detect structural similarities in protein whole structure [36, 37, 38], substructure [39, 40, 41], molecular surfaces [42, 43] et al., to infer similar protein function.

1.3.1 Specific Problems Studied in This Thesis

Within the broader classes of function prediction methods, this thesis studies one specific aspect of molecular function, called *binding specificity*, as a subproblem.

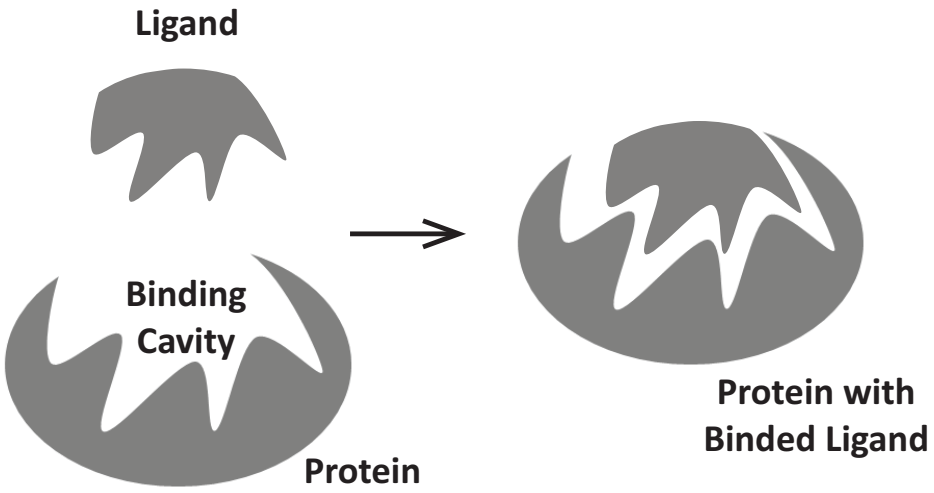

Figure 1.1: An illustration of protein ligand binding.

The thesis focuses on geometric comparisons to predict binding specificity. These comparisons come from two governing assumptions. First, different geometries of protein structures imply different specificities. Second, different geometries of protein electrostatics imply different specificities.

Proteins usually perform their biochemical function by attaching or binding to other molecules. While there are thousands of unique molecules, most proteins only attach to very specific binding partners. This property, of preferentially forming interactions with selective molecules, is called binding specificity. Building an understanding of the mechanism that achieves specificity is a common goal in many areas of molecular biology because it could reveal how teams of molecules function and how they might be manipulated or reengineered for medical purposes. Investigators, examine how mutations change specificities of cancer proteins to achieve drug resistance or produce artificial antibodies that selectively attach proteins of exotic bacteria to improve human immune systems. In this thesis, we study interactions between proteins and small molecules (*ligand*) as shown in Figure 1.1. The region where the ligand binds a protein is called the *binding cavity* or the *binding site*. Many observations show that binding cavities with similar geometries may be essential for accommodating identical ligand, while subtle structural variations in binding cavities could cause different binding specificities [43, 44].

1.3.2 Proteins are not Rigid Molecules

Proteins are generally thought to adopt unique structures that are determined by their amino acid sequences [45] and, as we will describe in the related works section 4.2.1, protein structures are usually taken as rigid objects in structure comparisons for function prediction. However, proteins are constantly in motion and are not strictly static molecules, but rather populate ensembles of conformations. In such cases, flexibility could affect protein structures at multiple levels: from local structures in individual atoms and amino acids [46], regional structure in intra-domains and multiple amino acid coupling [47] to global structures in multiple domains [48]. Studying protein structure flexibility is significantly important because it has been directly linked to functionally relevant phenomena such as allosteric signalling [48] and enzyme catalysis [49]. This thesis focuses on small structural motion within the binding cavity and design novel algorithms to analyze these motions to predict binding specificity. It is noted that structural changes resulting from these motions do not change specificity, but they create sources of errors for prediction methods.

With conformational flexibility of protein structures being considered, the input for specificity prediction is not only a protein $\mathbf{A}$ but all its conformations. Depending on how conformations of the same protein can be understood, two specificity prediction problems are investigated. First, all conformations of the same protein can be regarded as one unit of input. The specific problem to be addressed in this context is: we compare different proteins to predict their binding specificities where all conformations of the same protein must be considered, and we call it *aggregate prediction*. The problem of aggregate prediction can be formulated as follows:

- **Input**: A protein structure $\mathbf{A}$ and all its conformations $\{\boldsymbol{A_1}, \boldsymbol{A_2}, \ldots, \boldsymbol{A_N}\}$.
- **Output**: Predictive label of binding specificity on $\mathbf{A}$.

Second, each conformation snapshot of a given protein can be taken as an independent source of input and we compare all conformations of different proteins to predict specificity on each conformation. We call this problem *individual prediction*.

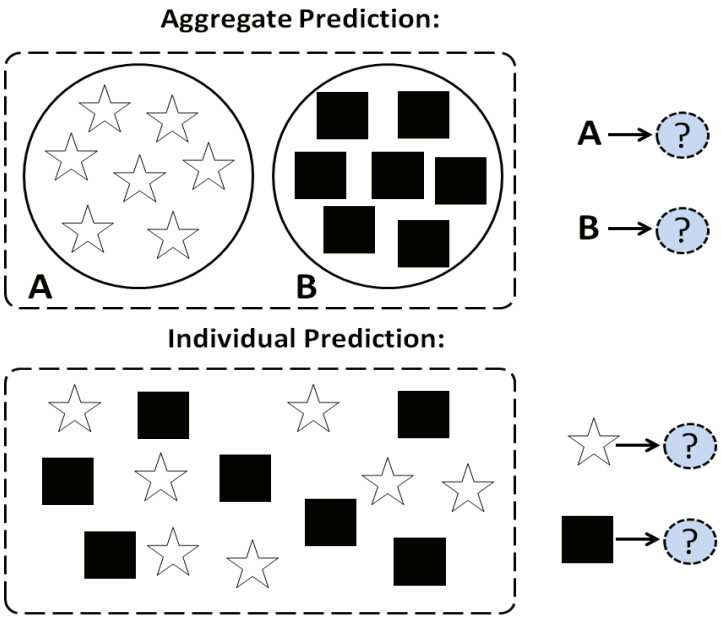

Figure 1.2: Illustration of definition of two specificity prediction problems. The star symbols represent conformation structures of the first input protein and diamond symbols represent conformation structures of the second input protein. The question marks represent the output information about the predicted binding specificity.

The individual prediction can be formulated as follows:

- **Input**: A protein structure **A** and all its conformations $\{\boldsymbol{A_1}, \boldsymbol{A_2}, \ldots, \boldsymbol{A_N}\}$.
- **Output**: Predictive label of binding specificity on each conformation $\boldsymbol{A_i}$.

These two problems are further illustrated in Figure 1.2. It is emphasized that these two problems are essentially different from existing function prediction because we leverage protein conformational flexibility to predict binding specificity.

1.4 Contributions

This thesis introduces several fundamental approaches to solve the aggregate prediction and the individual prediction. Instead of treating protein structures as rigid or partially rigid objects, our approaches leverage structural diversities in simulated protein conformations to predict binding specificity. We evaluate these approaches with application on two protein superfamilies.

1.4.1 Methods for Aggregate Prediction

The Flexible Aggregate Volumetric Analysis (FAVA) is the first conformationally robust tool for comparing proteins with similar binding cavities but different binding preferences. FAVA examines a large number of conformational samples to characterize local flexibility within the binding cavity. In particular, FAVA is capable of identifying the frequently conserved regions which is essential for accommodating the binding ligand and causing steric hindrance. FAVA provides an unsupervised learning tool that allows for the automatic detection of subfamilies with different binding specificities. FAVA also detect influential amino acids associated with differences in binding, predicting established experimental results.

Point-based Ensemble for Aggregate Prediction (PEAP) is a novel approach to enhance binding specificity prediction. This method identifies protein substructure matches to extract atomic positions of influential amino acids across all protein conformations. This capacity provides a novel representation of molecular flexibility in the binding cavity using conformational samples. Additionally, our method employs ensemble clustering techniques to incorporate the diversity of structural motions in the binding cavity, which helps to mitigate prediction errors. Although there are several works that apply ensembles of machine learning classifiers to predict protein function [50, 51, 52], our method is the first ensemble-based procedure to predict binding specificity in an unsupervised way.

1.4.2 Methods for Individual Prediction

The first individual prediction method is an atomic point representation. This is the first known to analyze maps of binding cavity conformations in order to perform an unsupervised specificity prediction. This representation detects coordinates of selected amino acids to represent the binding cavity of a protein conformation as a high dimensional point. Effective dimension reduction methods map each point in a lower embedding feature space. This map visualizes a high level organization of binding cavity conformations and makes specificity prediction with high accuracy.

The second individual prediction model is a volumetric lattice representation. Different from the atomic point representation that only describes Carbon alpha atoms, this model enables an all-atom motion representation of the binding site. Furthermore, the volumetric lattice representation localizes the binding cavity into many tiny user-defined cubes for feature extraction. This representation proves to be more informative than traditional point-based representations for binding cavity comparisons, and could be a general tool for protein structure analysis.

In the end, we introduce an electrostatic lattice model fore specificity representation. This representation builds a lattice model on protein electrostatic isopotentials to compute discriminative features. By ignoring atomic points or molecular surfaces, this representation provides the first method to predict binding specificity which is independent of protein structure comparisons.

1.5 Thesis Schedule

Chapter 2 includes a summary of protein structure comparisons. Chapters also discusses why molecular dynamics simulation and electrostatic potentials are used in this thesis. Chapter 3 introduces the data sets and demonstrates the considerable structural variations in the binding site. Chapter 4 introduces two methods for aggregate prediction: FAVA and PEAP. Chapter 5 introduces methods for individual prediction: an atomic point representation, a volumetric lattice representation and an electrostatic lattice representation. Chapter 6 summarizes the thesis and proposes several future works.

Chapter 2

Related Works

Ligand binding specificity is a property that is fundamentally defined by comparisons: A protein prefers one ligand because it binds with less affinity to other ligands, not because of absolute affinity. These preferences arise because the preferred ligand has a geometric shape or an electrostatic distribution that is more complementary to the protein than other ligands. The comparative nature of binding preference makes protein structure comparison an ideal class of methods for examining specificity.

We begin with describing existing approaches for protein structure comparisons. One challenge for these comparative methods is the assumption of treating protein structures as rigid or partially rigid objects, but they cannot fully represent flexibilities of protein structures. For this reason, we then discuss molecular dynamics simulation, which is a more comprehensive way for flexibility representation. Finally, we introduce protein electrostatic potentials because electrostatic charges could selectively influence ligand binding.

2.1 Protein Structure Comparisons

2.1.1 Rigid Structure Comparison

The comparison of protein structures depends highly on how they are represented. Most comparison approaches take the *rigidity assumption* that protein structures

are treated as rigid objects. These methods can be further categorized into three types: point-based representation, surface-based representation and learning-based representation.

Point-based representations extract coordinates of atoms or amino acids and estimate mappings between the function of a protein and atomic points it contain. For example, DALI [53], transforms an input protein into a matrix of distances between all their Carbon alpha atoms, and overlaps along the matrix diagonal indicates similarity in adjacent fragments. DALI builds sets of submatrices of fixed size by segmenting the original matrix into regions of overlap. In such a representation, submatrix matches can be assembled into a final superposition between two protein structures. The optimal superposition can be obtained using the branch and bound algorithm Other representative methods model protein whole structures using points in three dimensional space [54, 55, 37, 56, 57, 41] or nodes in geometric graphs [58, 59]. A second type of point-based representation encodes atom positions in protein substructures, such as amino acids in functional sites or binding sites [60, 61, 62], evolutionarily influential amino acids [63] or pseudo atoms [64]. These selected amino acids or atoms can be generated by expert manual selection [39, 63], literature search [65] or database retrieval [66, 67, 68] and they can be further refined for better functional annotation [69, 70, 71, 72, 73]. Point-based methods achieve extreme efficiency, and is capable of searching proteins with similar structures from a large dataset.

Point-based representation takes the assumption that molecular interactions that affect protein function can be traced back to the position of atoms or amino acids. However, in many case studies, ligand binding occurs due to complementarity of molecular surface. How proteins interact with other molecules highly depends on the shape of surface clefts or binding cavities because they provide more space to form complementary hydrogen bonds, hydrophobic contacts or electrostatic interactions [74, 75]. All observations inspire the design of surface-based representation. Surface-based methods employ closed surfaces or surface patches to approximate solvent-

accessible shape of protein clefts or cavities for accommodating solvent within the binding cavity. The approximation can be built using triangular meshes [76, 77, 43, 78], alpha shapes [42, 79], three dimensional grids [80] and spherical harmonics [81, 82]. Surface-based methods are essentially specification of protein structure using a finer resolution than that of selected atoms or amino acids and protein structure comparisons are based on matching patterns between solvent-accessible surfaces.

Protein structures can also be represented as feature vectors because of the natural mapping of function prediction to building machine learning models. The first type of learning-based methods explicitly calculate each feature attribute that can be extracted from protein structures [83, 84, 85] or structural datasets [86]. For example, SCREEN [84] collected a list of 408 structural and physiochemical properties, such as number of atoms, maximum depth and average curvature, of the binding cavity with application on drug-binding prediction. Instead of computing protein features explicitly, the second type of learning-based methods describe protein structures with higher level of abstraction, such as random work on graphs [87], structural kernels on user-defined geometric shapes [88, 40] and pairwise similarity via structure matches [89, 90, 70]. Due to the exponentially increasing number of protein structures, comparisons that rely on machine learning has attracted more attention and has been largely applied to the recent large-scale protein function prediction assessment [91]. It is also noted that methods introduced in this thesis generally belong to learning-based approaches because of feature representation of protein structures.

2.1.2 Flexible Structure Comparison

One issue of protein structure comparisons is that, in many cases, the rigidity assumption cannot be taken for granted. For example, conformational changes in structures of exotic proteins could cause drug resistance and how these flexibilities should be represented is the key to manipulate the specificity we desire. In the

meanwhile, without rigid simplification, structure comparisons could become more difficult. First, structural flexibility creates comparison errors because conformations of the same protein may get falsely recognized as different proteins. Second, comparative analysis requires more computational efforts where all flexible structure elements must get considered. Many works reported comparisons that tolerate flexible representation of protein structures. Ye and Godzik developed a flexible structure comparison tool called FATCAT [92]. FATCAT is different from rigid comparisons because it adds a limited number of structural twists between aligned protein fragments which are taken as rigid bodies. The final similarity between two proteins includes the score between aligned fragments and alignment penalty of twists that connect the fragments. Other flexible comparisons employ hinges [93, 94], graphs [95, 96] or dynamics programming [97, 98, 99] to encode protein structures as rigid substructures that are joint by flexible linkers.

However, these representations are essentially partially rigid representations, and they cannot examine every flexible element that influences specificity. For example, rigid or partially rigid representations do not examine small motions in backbone atoms or motions in sidechain atoms, but these structural motions could change the shape of the binding cavity and affect specificity [43].

2.2 Molecular Dynamics

The previous section reveals the limitations of existing methods for flexible structure comparisons. It is clear that a more comprehensive way for flexibility representation is to treat every protein atom as a flexible component. Therefore, we will introduce *molecular dynamics* (MD), a computational simulation system for studying physical movements of molecules. In this thesis, MD simulation generates inputs for our trainable methods that accommodate for motion in every protein atom.

Using molecule snapshots generated by MD simulations as a flexible structure representation, enjoys several advantages. First, MD simulations generate the time-dependent trajectory of every atom, which is a more general description of steric

movement of every molecular component. Coordinates of every atom may produce functionally related elements that could be overlooked by rigid or partially rigid representations. Second, protein conformations generated by MD simulations are constrained by biophysical properties. Therefore, every snapshot is a semi-realistic conformation, which cannot be guaranteed by existing representations because they only consider motions of fragment linkers. Figure 2.1 shows the structural flexibility of the pseudomonas mandelate racemase protein using conformational samples that are generated by MD simulations. It reveals that the motion of every atom is represented using selected conformations.

In the classical molecular mechanics molecular dynamics (MM-MD) simulation, the state of every atom, e.i. the position, the velocity and the acceleration, is determined by the Newtonian laws of motion equation. The forces and potential energies between all atoms are computed using molecular mechanics force fields, including spring forces capturing bonds, angles and dihedral angles, Van der Waals interactions and electrostatic Coulomb contributions. Given initial conditions, MM-MD simulation iteratively moves the system to a local optimal by minimizing the energy landscape in finite steps where derivative based optimizations, such as the steepest descent and the conjugate gradient, are often applied.

Besides MM-MD, many other types of MD have been proposed depending on different systematic considerations. In coarse-grained (CG) models [100], the molecule structure can be alternatively represented into a simplified form. For example, adjacent amino acids may move sufficiently in concert to be encoded as a single point. CG simulations are much more computationally fast and are suitable for modelling dynamics of large-scale molecular systems. Different from MM-MD that is based on Newton equations, the quantum mechanics molecular dynamics (QM-MD) studies simulation using Schrödinger equations where an electron is described as a unique wave function for describing the quantum state. Although QM-MD becomes much more computationally intensive, the increase of accuracy in molecular modelling usually requires its application in simulating key parts of a protein [101]. Instead of

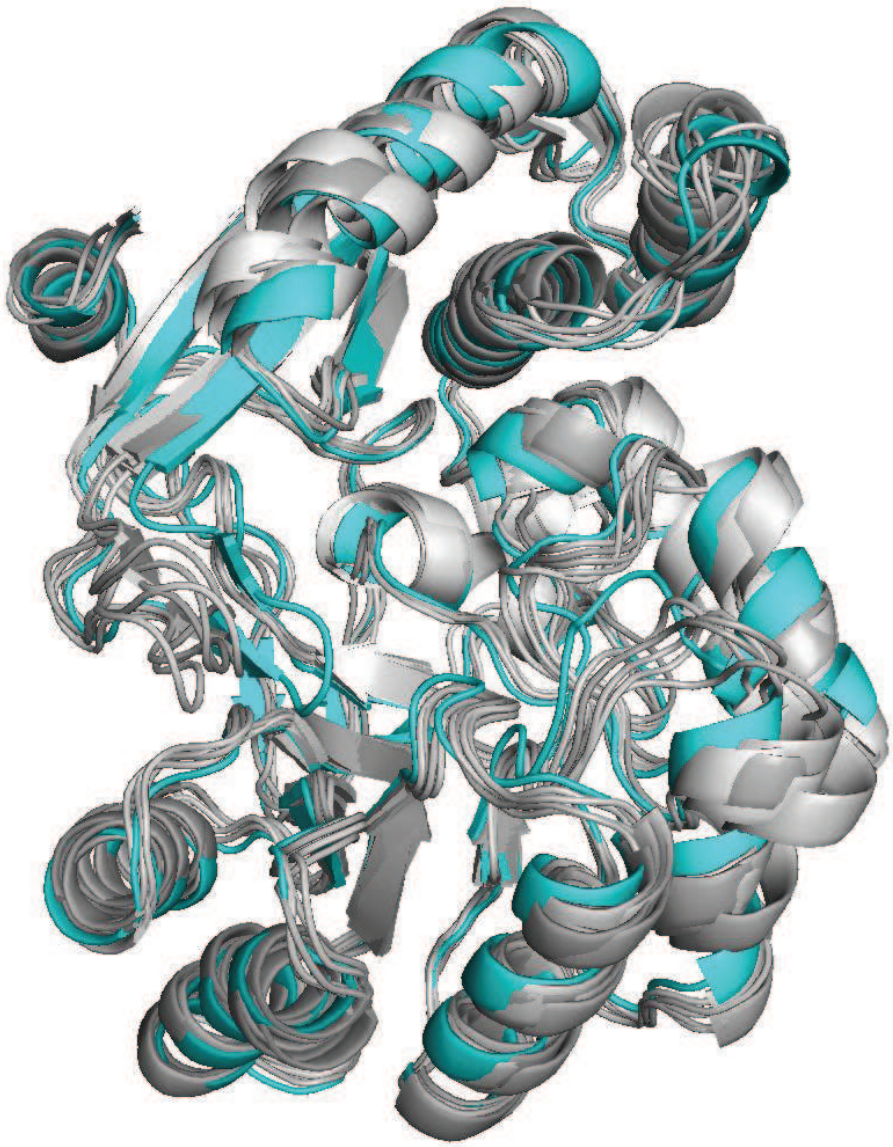

Figure 2.1: Conformational samples (grey) of the whole structure of pseudomonas mandelate racemase (pdb: 1mdr) with respect to its original structure (teal).

relying on physical laws, a system can alternatively utilize Monte Carlo simulation to randomly move atoms and accept resulting atoms by examining the associated energy. Monte Carlo MD is less computationally costly but the simulation could become too flexible due to its stochastic nature [102].

In this thesis, we apply the classical MM-MD to simulate protein structures for flexible representation. In Chapter 3.3, we will present more technical details of our simulations.

2.3 Electrostatic Potentials

Protein comparisons reported above use either atomic positions or molecular surfaces to compare protein structures. They exploit the facts that protein structures define patterns of steric hindrance that is imposed on potential binding partners, so they are logical choice for comparative analysis. Nonetheless, when considering all influences on molecular recognition, longer distance electrostatic effects can have a selective influence on binding partners even before they come into contact with molecular surfaces. In such cases, a comparison of electrostatic potentials may reveal complementary information for binding preferences that may not be encoded in structure comparisons. An example of electrostatic effects is shown in Figure 2.2.

Protein electrostatic potential, the summation over all atom charges, is the potential energy of a proton at a particular location near the protein surface. Negative electrostatic potentials represent the attraction of the proton by the concentrated electron density while positive potentials represent repulsion of the proton. The unit of electrostatic potential is kT/e where k is Boltzmann constant, T is temperature in Kelvin and e is the charge of an electron. In practice, electrostatic potentials can be computed by solving the Poisson-Boltzmann equation [103].

Several efforts have been made to analyze molecular electrostatic potentials that reveal protein function. Some quantified charge distributions over whole protein structures [104, 105] or local regions such as protein domains [106], active sites [107] protein-protein binding interfaces [108] or structural motifs [109]. Few more methods

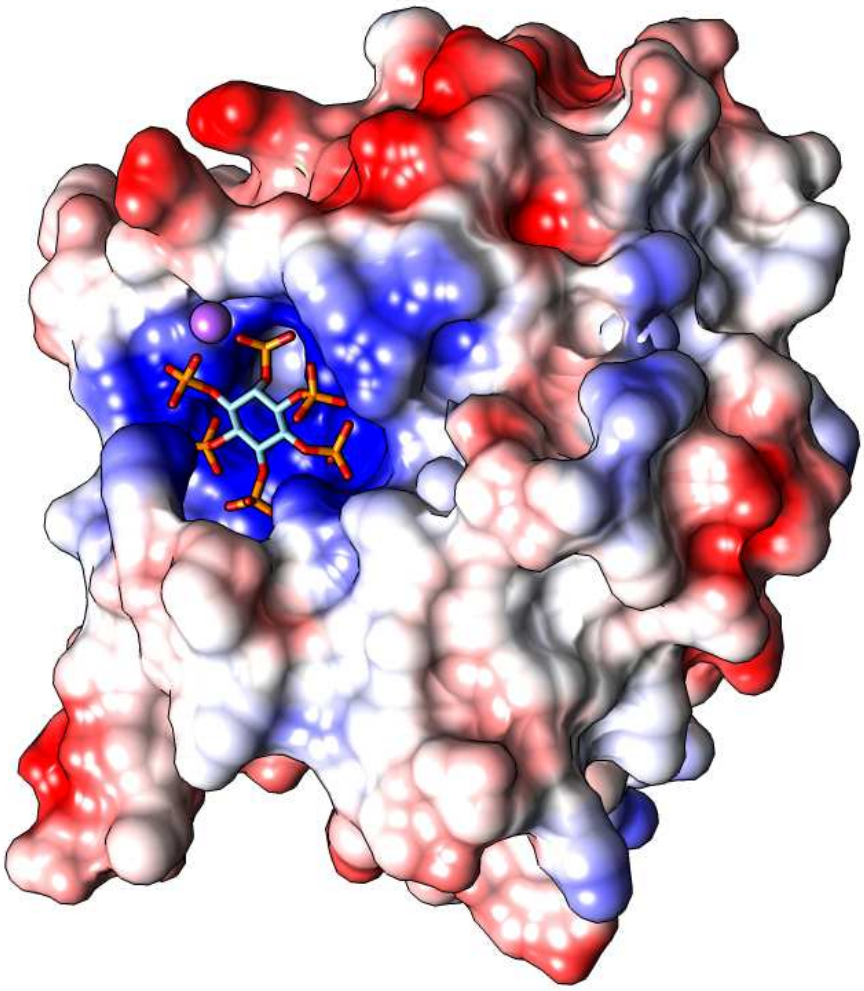

Figure 2.2: The molecular surface of a Vibrio cholerae RTX cysteine protease (pdb:3*eeb*) where the electrostatic potential energy was mapped onto the surface. The area of the binding cavity is strongly positively charged (blue) which is surrounded by areas of neutral charge (white) and areas of negative charge (red). The inositol-hexakisphosphate (IHP) ligand, which is strongly negatively charged and attracted by the positive binding cavity of 3*eeb*, is shown in sticks.

compared electrostatic potentials directly by computing a similarity index [110] or constructing tree-based structures [111]. Kinoshita et al. compared electrostatic potentials on molecular surfaces to infer protein function [112, 113]. However, these methods did not focus on ligand binding specificity. To deal with this issue, in section 5.3, we will discuss a structure-independent approach to predict specificity using protein electrostatic potentials.

Chapter 3

Datasets

This chapter describes the data sets used in the following chapters. The data sets consist of two nonredundant enzyme superfamilies: serine protease superfamily and the enolase superfamily. These two sets were selected based on established results recording the existence of distinct subfamilies in each superfamily with different binding preferences. Within serine proteases, we selected trypsin, chymotrypsin, and elastase subfamilies. In the enolase superfamily, we selected enolase, mandelate racemase, and muconate lactonizing enzyme subfamilies.

3.1 Protein Family Selection

The serine proteases hydrolyze peptide bonds by recognizing a set of adjacent amino acids with specificity subsites that are numbered $S_4, S_3, S_2, S_1, S'_1, S'_2, S'_3, S'_4$. Each subsite has binding preferences on one amino acid before or after the $S_1 - S'_1$ hydrolyzed bond. In this work, we focus on three different binding specificities of the S_1 subsite: positively charged amino acid [114] for trypsins, large and hydrophobic amino acid [115] for chymotrypsins and small hydrophobics [116] for elastases.

The enolase superfamily proteins catalyze reactions by abstracting a proton from a carbon adjacent to a carboxylic acid [117] near the C-terminal domain of beta sheets of the conserved TIM-barrel structures. In this work, we study three different catalysts. The enolase subfamily converts 2-phosphoglycerate (2-PG) to

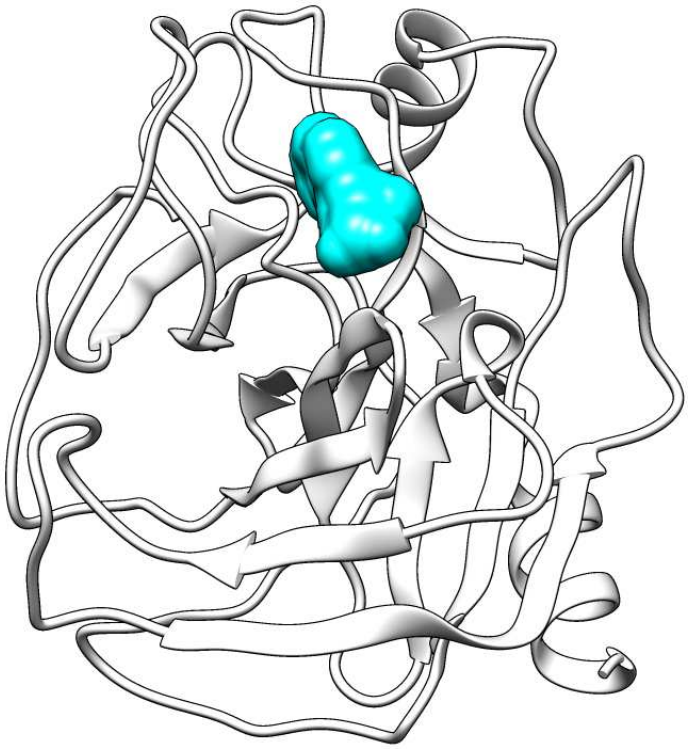

Figure 3.1: The crystal structure of a cold-adapted fish species trypsin (pdb:1a0j) is shown above. The binding ligand is shown in red surface representation. This figure is generated with UCSF Chimera [1].

phosphoenolpyruvate (PEP) [118], the mandelate racemases convert between (S)-mandelate and (R)-mandelate [119] and the muconate-lactonizing enzymes convert lignin-derived aromatics, catechol and protocatechuate to citric acid cycle intermediates [117].

For example, the protein structure of the Atlantic salmon trypsin (pdb:1a0j) (Figure 3.1) has been labelled as EC class 3.4.21.4. The definition of each component is:

- EC 3 enzymes are hydrolases
- EC 3.4 are hydrolases acting on peptide bonds

- EC 3.4.21 are hydrolases that cleave peptide bonds in which serine serves as the nucleophilic amino acid
- EC 3.4.21.4 are those that cleave peptide chains mainly at the carboxyl side of the amino acids lysine or arginine

The EC classification for serine proteases and the enolases is shown below:

Protein Superfamily	EC Number	Enzyme Family
Serine Proteases	3.4.21.1	Chymotrypsins
	3.4.21.36	Elastases
	3.4.21.4	Trypsins
The Enolases	4.2.1.11	Enolases
	5.1.2.2	Mandelate Racemases
	5.5.1.1	Muconate Lactonizing Enzymes

Table 3.1: EC number used in the data set

Our data sets contain families of proteins with highly similar enzymatic function (e.g. 3.4.21.1 and 3.4.21.4) and families of proteins that differ in function at all 4 levels (e.g. 4.2.1.11 and 5.1.2.2).

3.2 Protein Structure Selection

Serine protease and enolase structures were selected from the protein data bank (PDB) [35] on 6.21.2011. Based on enzyme classifications (EC), the PDB contained 676 serine proteases and 66 enolases in the families selected for our data set. From these structures, mutants, structures with disordered regions, and enolases with closed or partially closed capping domains were removed. Next, one structure from

Serine Protease Superfamily:
Chymotrypsins: 1ex3
Elastases: 1b0e, 1elt
Trypsins: 1a0j, 1ane, 1aq7, 1bzx, 1fn8, 1h4w, 1trn, 2eek, 2f91
Enolase Superfamily:
Enolases: 1ebh, 1iyx, 1te6, 3otr
Mandelate Racemase: 1mdr, 2ox4
Muconate Lactonizing Enzyme: 2pgw

Figure 3.2: PDB codes used in the data set.

any pair of structures with greater than 90% sequence identity was removed, with a preference for keeping structures associated with publications. Technical problems with simulation prevented proteins 8gch, 1aks, and 2zad from being added. From the 12 serine protease and 7 enolase structures that remained, ions, waters, and other non-protein atoms were removed. Hydrogens, unavailable in some structures, were also removed, but non-canonical amino acids (e.g. selenomethionines) were not removed. All the structures are shown in Figure 3.2 by their PDB code and are classified into subfamilies by their binding specificities.

Alignment. We superposed all structures and conformational samples in each superfamily. All serine proteases and their samples were superposed onto 8gch, a chymotrypsin, and all enolases and enolase samples were superposed onto 1mdr using ska [57]. These structures were selected because of the presence of a bound ligand, which we used to define the binding cavity.

3.3 Protein Structure Simulation

The conformational samples of each protein structure were simulated using GROMACS 4.5.4 [120]. The input structure was centered inside a cubic waterbox using a 3-point solvent model SPC/E [121]. The waterbox was set so that there is at least 10 Å between the protein and the nearest part of the box. Charge balanced sodium and potassium were then added at a low concentration ($< 0.1\%$ salinity). Steepest descent was used to minimize energy on the entire simulation system. Isothermal-Isobaric (NPT) equilibration was performed in four 250 picoseconds steps for temperature and pressure equilibration before the primary simulation. Over the four 250 picosecond minimization period, at 1000 $kJ/(mol * nm)$, each equilibration step reduced the position restraint force by 250 $kJ/(mol * nm)$. Backbone positions constraints were released during the NPT simulation and system energies were computed in the beginning of the equilibration phase. Temperature was set to 300 Kelvin and pressure was set to 1 bar. Temperature coupling was computed using Nosé-Hoover thermostat [121] and pressure coupling was computed using the

Parrinello-Rahman algorithm [122, 123]. The simulation used P-LINCS [124] to update bonds and used particle mesh Ewald summation (PME) [120] to calculate electrostatic interaction energies. The primary MD simulation was started using the atomic positions and velocities of the final equilibrium state.

The simulation was maintained for 100 nanoseconds with 1 femtosecond timesteps on multiple 16 core nodes of the Lehigh corona server where OpenMPI was used for parallel communications. The trajectory file was convert to the PDB format with only atomic positions. 600 samples were selected of each protein structure at uniform intervals.

3.4 Binding Cavities Vary Considerably

As shown in Figure 2.1 as an example, conformations of the same protein used in our data set exhibit identical folds and highly similar whole structures. Nevertheless, the globally structural similarity does not indicate the similarity in the binding cavity and we will show that binding cavities in both serine proteases and the enolases vary considerably.

Figure 3.3 exhibits structural variations in selected conformational samples of pseudomonas mandelate racemase (pdb:1mdr) from the 100 nanosecond simulation as described earlier. It is observed that smaller backbone motions and side chain motions could shrink, enlarge or separate the binding cavity. The pseudomonas mandelate racemase is no the most flexible structure in our dataset, and similar motions in the binding cavity can be found in almost all other structures in our data sets.

In Figure 3.4, we plot the volumes of binding cavities in conformational samples of the entire dataset. Among the serine proteases, samples of trypsin cavities ranged from 248 $Å^3$ to 692 $Å^3$, chymotrypsin cavities ranged from 276 $Å^3$ to 568 $Å^3$, and elastase cavities ranged from 126 $Å^3$ to 552 $Å^3$, despite the general principle that chymotrypsin $S1$ cavities are larger to accommodate aromatic sidechains, and elastase cavities are smaller to accommodate amino acids like alanine or valine.

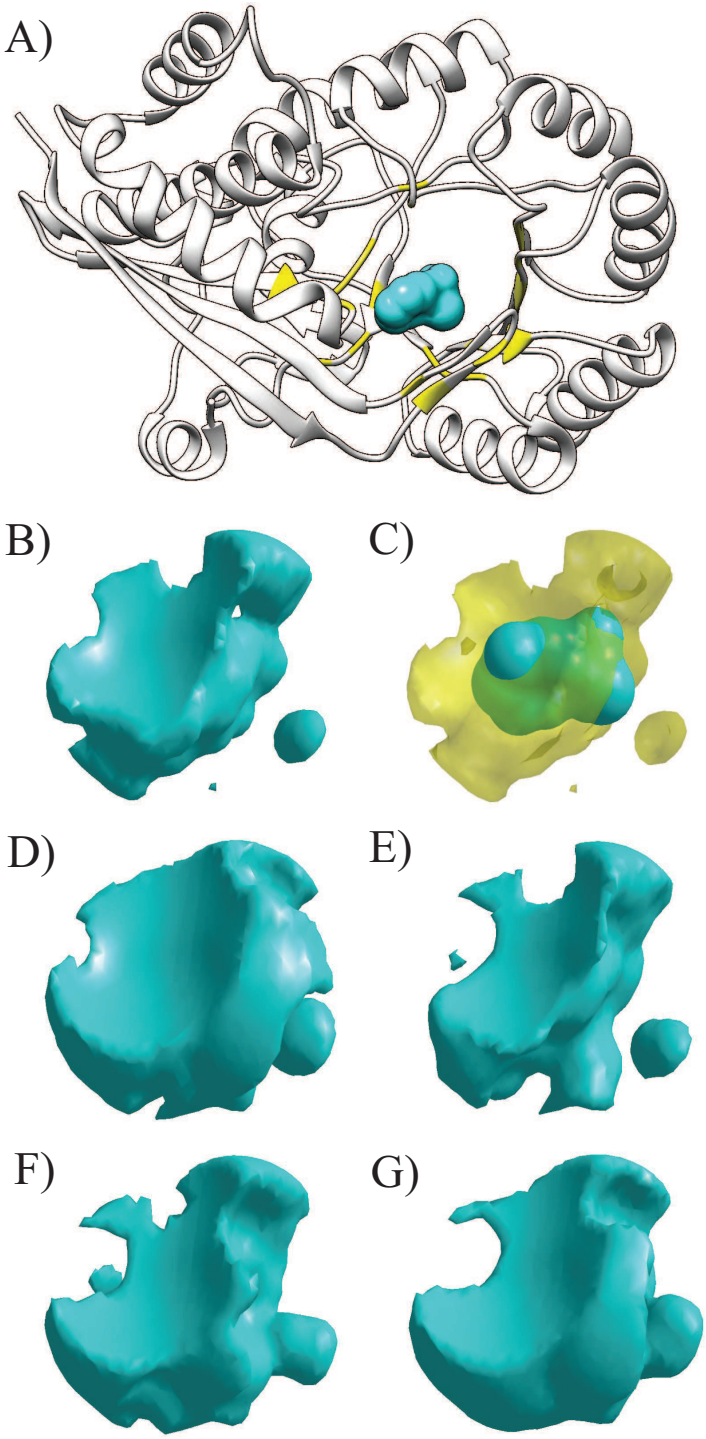

Figure 3.3: Conformational samples of the binding cavity in pseudomonas mandelate racemase (pdb: 1*mdr*). A) The position of the binding ligand (teal) is mapped on to the tertiary structure of racemase protein (white) where top 20 amino acids that are nearest to the binding cavity is also visualized (yellow). B) The binding cavity in the naive crystal structure. C) The binding ligand (teal) within the same binding cavity (transparent) in B). D-G) Binding cavities from selected conformational samples that are generated by MD simulations. All these cavities are rendered from the same perspective.

Similar variations can be seen amongst the ligand binding cavities of the enolase superfamily. Enolase cavities ranged from 90 $Å^3$ to 507 $Å^3$, mandelate racemases ranged from 225 $Å^3$ to 673 $Å^3$, and cavities sampled from muconate lactonizing enzyme were between 89 $Å^3$ and 343 $Å^3$. This degree of structural variation demonstrates the fundamental difficulty of accurately comparing binding site geometry in the presence of flexibility.

Statistical modelling with a rigid model for classification does not add precision to the structural comparison of flexible binding sites. We used VASP-S [44], a statistical modelling tool that isolates structural element between protein cavities that may influence preferential binding, to generate structural fragments between pairs of cavities sampled from serine proteases. It was observed that more than 65 percent of all fragments were incorrectly classified as being so large as to be consistent with different binding preferences, and it suggests that a comparison of individual structures has a high probability of being inaccurate.

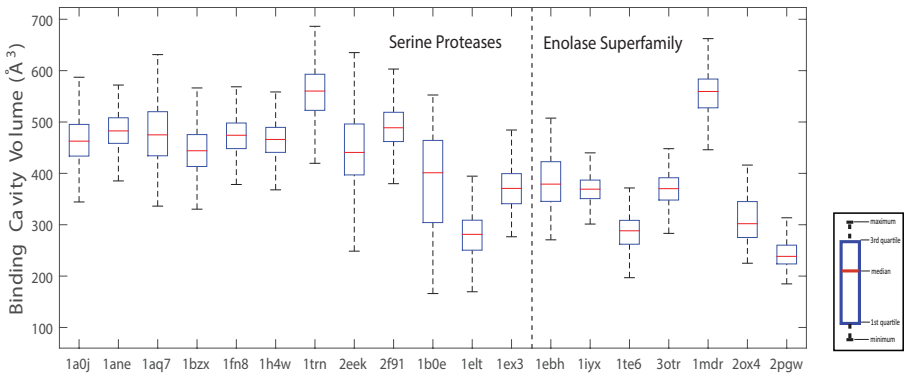

Figure 3.4: Aggregate variations in cavity volume in our whole data set. Cavity of almost all proteins varied considerably.

From these observations, it is clear that the flexibility of serine proteases and the enolases creates significant variations between different samples of binding cavities from the same protein. Because of the variability in the data, flexible protein comparisons within the binding cavity, which is the specific problem studied in this thesis, is not a trivial problem. Thus, techniques that will be introduced in the following two chapters, which incorporate flexibility from conformational samples

into the analysis, are essential for accurate general comparison.

Chapter 4

Aggregate Prediction Pipelines Development

In this chapter, we will introduce two fundamental methods for the aggregate prediction. The input is a superfamily of protein structures with different specificities where each structure is presented with a set of conformational samples and the aggregate prediction outputs the predictive label on each input protein.

FAVA is a novel volumetric method for geometric comparisons of similar binding cavities with different specificities. FAVA integrates randomly selected samples into a three dimensional conserved region of the binding cavity as an aggregate representation and hierarchical clustering is then applied to categorize proteins to predict specificity. However, the conserved region does not necessarily exist in highly flexible binding sites. To enhance specificity prediction, a second method PEAP randomly selects one conformational sample from each protein for generating a base prediction, and ensemble clustering integrates base predictions from many samplings for a consensus prediction.

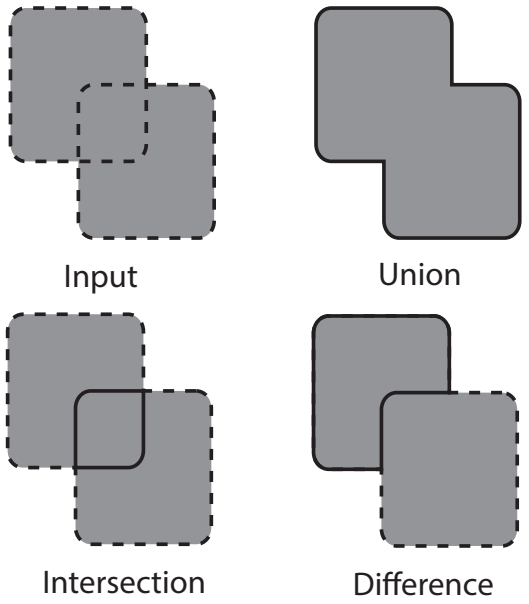

Figure 4.1: The CSG operations used by VASP, with input regions (light grey, dotted outline) and output regions (solid outline).

4.1 FAVA: A Volumetric Method for Flexible Protein Structure Comparisons

4.1.1 Method Overview

In section 3.4, it was observed that binding cavities in both serine proteases and the enolases vary considerably. To mitigate comparison errors caused by flexibilities of the binding cavity, our approach with FAVA is to represent the a conserved region that is frequently, but not universally, within the ligand binding cavity. We call this region the *frequent region* because it ignores the geometry of unusual conformations that can obfuscate the solvent-accessible region. Below, we first describe how compute frequent regions using a series of Constructive Solid Geometry (CSG) [125] operations, such as union, intersection and differences (Figure 4.1) in which the description of each individual CSG operation can be found in [43]. We then describe how we compare frequent regions from multiple proteins to identify conserved and varying space within frequent regions for predicting binding specificity. Finally, we explain how we use solid representations to characterize the flexibility of individual

amino acids and their steric impingement on nearby binding cavities.

4.1.2 Generating Frequent Regions

Solid Representation of Binding Cavities

As input, we require the overlap threshold k, N conformational samples of a protein structure A, and a ligand l bound to A. We refer to the conformational samples as $\{A_0, A_1, \ldots A_N\}$. First, every sample A_i is superposed onto A by minimizing the root mean squared distance between identical amino acids. Next, in every A_i, we use GRASP2 [126] to generate the molecular surface $m(A_i)$. This surface is defined by the classical rolling probe algorithm [127] with the standard probe size of 1.4Å. Since every conformational sample is superposed onto A, we use l to locate the ligand binding site in every superposed $m(A_i)$.

Second, at every atom in l, we center a sphere with radius 5 Å . The CSG union of the spheres defines a neighbourhood, S_l, that defines the vicinity of the ligand binding cavity in every sample. Next, we generate the *envelope surface* $e(A_i)$ for every sample. This procedure is similar to molecular surface generation except that the envelope surface is calculated with a 5.0 Å radius probe. Since we are taking a much larger probe size, the envelope surface does not roll into small binding cavities, thereby making the envelope surface the boundary that separates the cavity from the solvent. To mitigate the cavity shape variations that caused by envelope surface differences from multiple samples, we compute intersection of all envelope surfaces, $E(A) = \bigcap_{\forall i} e(A_i)$, which is referred as the global envelope surface.

Finally, for every sample, we compute the intersection of the global envelope surface and what remains of ligand spheres in the molecular surface, $a_i = (S_l - m(A_i)) \cap E(A)$. a_i is a solid representation of the binding cavity on the sampled structure A_i. This solid cavity generation procedure is more detailed in [128]. The solid representation is selected because it describes geometries of all atoms that can sterically hinder binding.

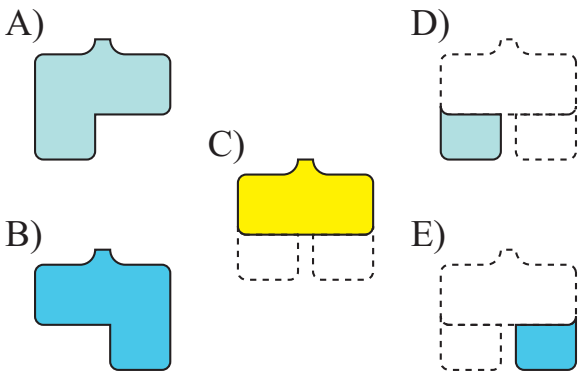

Figure 4.2: A comparison of frequent regions. A,B) Frequent regions $\alpha_k^\star$ (teal) and $\beta_k^\star$ (light blue). C) Conserved frequent region, $FC(A, B)$ (yellow). D,E) unconserved frequent regions (teal, light blue).

Solid Approximation of Frequent Regions

We use the sampled cavities a_i together to approximate the frequent region α_k. Before we approximate this region, it is critical to recognize first that computing α_k explicitly, on a protein with many sampled conformations, is computationally impractical for many k. Consider, for example, the simple case of $k = 30$. The region α_{30} includes the CSG intersection of $a_0, a_1, a_2, \ldots, a_{30}$, because any point inside all of these regions is inside at least 30 a_i, and thus inside α_{30}. The same is true for any thirty member subset of $\{a_0, a_1, \ldots, a_N\}$, so α_{30} is the union of all intersections of thirty distinct sample cavities: $\binom{N}{30}$ intersections. Where N is several hundred samples and k is nontrivial, the exponential size of the calculation is clearly impractical, given the number of combinations.

FAVA approximates α_k by randomly selecting subsets of size k. We call the approximated result $\alpha_k^\star$, and compute it in the following manner: given any k, we randomly select 500 distinct subsets of $\{a_0, a_1, \ldots, a_N\}$ of size k, and compute their CSG intersection. Finally, we compute the CSG union of the resulting intersections, $\alpha_k^\star$. While random selections of different sizes were tested, frequent regions based on different random subsets of 500 had consistent volumes. We deemed 500 samples to be sufficient for accurate representations.

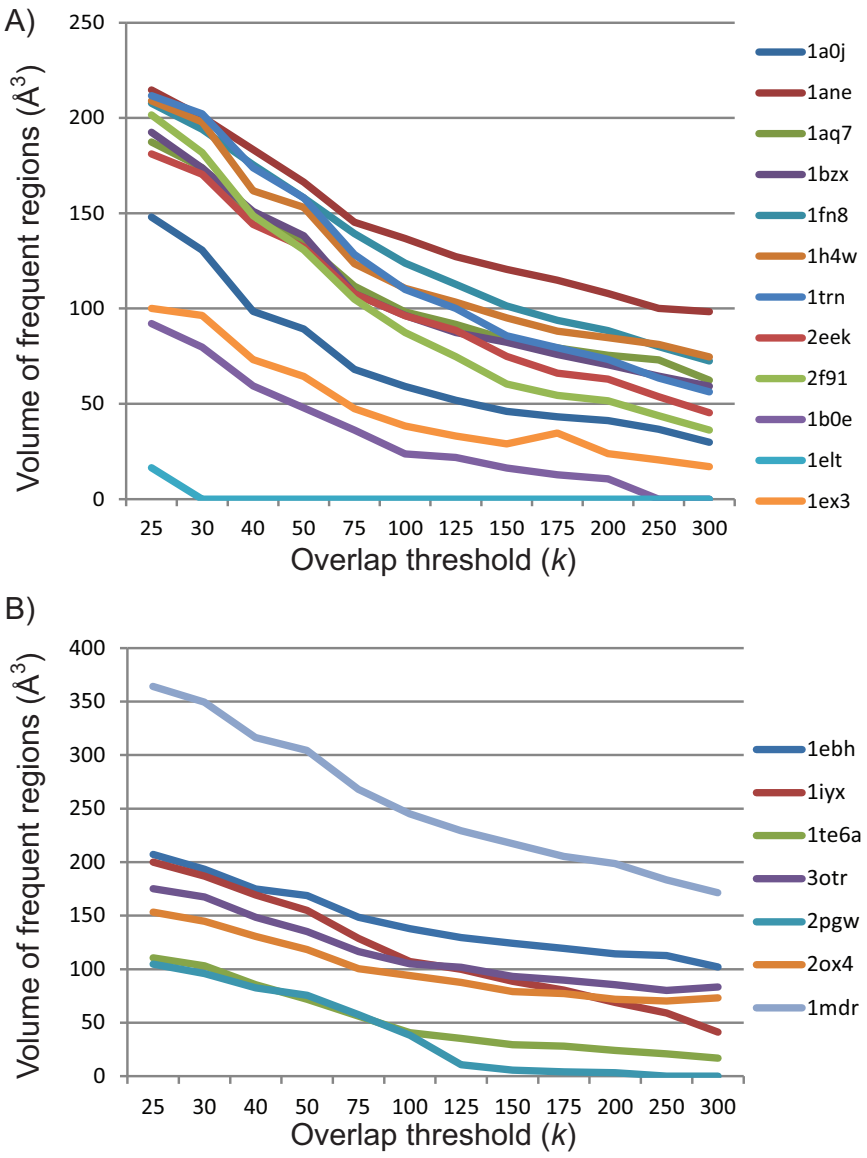

Figure 4.3: Volumes of frequent regions in serine protease (A) and enolase (B) cavities, computed at varying thresholds.

4.1.3 Evaluating Frequent Region Approximation

FAVA approximates frequent regions using random selections of conformational samples. Actual frequent regions cannot be computed on realistic data because of their combinatorial nature. This situation prevents a direct evaluation of the accuracy of our approximation technique, but it does not prevent us from evaluating the geometric consistency of the approximations generated. Specifically, when considering conformational samples from the same protein, frequent regions with higher overlap thresholds must always have equal or smaller volume than frequent regions with lower overlap thresholds. This fact holds logically because regions where k cavities overlap are also, by definition, a region where fewer than k cavities overlap.

We evaluated the degree to which this rule holds for our approximation by computing the volumes of frequent regions at a wide range of overlap thresholds for all proteins in our data set. Figure 4.3 indicates that volumes of frequent regions are almost monotonically descending as overlap thresholds increase. They also indicate that frequent regions from some proteins remain consistently larger than others, suggesting fewer conformational changes that interfere with the shape of the binding cavity. It is noted that, though not inconsistent, that volumes of frequent regions from sampled cavities of Atlantic salmon elastase (pdb: 1elt) become zero above overlap thresholds of 25, indicating that conformational flexibility radically alters the shape of that cavity. Overall, these observations suggest that FAVA is generating stable, logically consistent approximations of frequent regions.

4.1.4 Comparing Frequent Regions

Given two proteins A and B, we use their frequent regions $\alpha_k^\star$ and $\beta_k^\star$, to evaluate the similarities and differences of their ligand binding sites over time (Figure 4.2). These calculations are only performed once both structures and all conformational samples are structurally aligned, to avoid errors from poor superposition (e.g. poor registration). We use the frequent regions to identify conserved frequent regions.

The conserved frequent region between the samples of A and B is $\alpha_k^\star \cap \beta_k^\star$ (Figure.

4.2C). Because the conserved region is the space intersecting two frequent regions, it approximates a binding cavity region that is solvent accessible in both proteins in more than k conformational samples. We measure the *volumetric distance*, $D(A, B)$, between the frequent regions of two proteins using the following expression:

$$D(A, B) = 1 - \frac{|\alpha_k^\star \cap \beta_k^\star|}{|\alpha_k^\star \cup \beta_k^\star|}, \tag{4.1}$$

where the expression $|x|$ denotes the volume within a solid region x.

To compute the volume of the a $3D$ solid region represented by a boundary surface that is composed of oriented triangles, we first calculate the centroid c of all triangle corners. For every triangle, we compute the centroid t_c and the normal vector t_n. We then decide if the selected triangle faces away from c or towards c by measure the dot product between t_n and $t_c - c$. Next, we connect three corners of the triangle and c to generate the tetrahedron. We add the tetrahedron volume if the triangle faces towards the global centroid c and subtract the tetrahedron volume otherwise. The volume of a tetrahedron can be evaluated by Tartaglia's Rule [129].

To evaluate FAVA, we generated frequent regions with an overlap threshold of 50, and measured volumetric distance between all pairs of frequent regions in the same superfamily. Given a protein superfamily $F = \{f_1, f_2, ..., f_m\}$, the volumetric feature $\mathbf{v}_i$ for protein f_i is a vector defined as $\mathbf{v}_i = \{D(f_i, f_1), ..., D(f_i, f_m)\}$. The feature space of all-against-all volumetric distance within the protein superfamily can be represented by a matrix $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_m\}$. We then perform UPGMA clustering (Unweighted Pair Group Method with Arithmetic mean) [130], an agglomerative hierarchical clustering method with average linkage, to generate clustering based on the feature matrix.

Since frequent regions avoid inaccuracies that may be derived from individual conformational samples, we compared frequent region clustering against 10 clusterings of individual binding cavities from conformational samples selected randomly from each simulation. All clustering hierarchies were visualized using Newick Utilities [131].

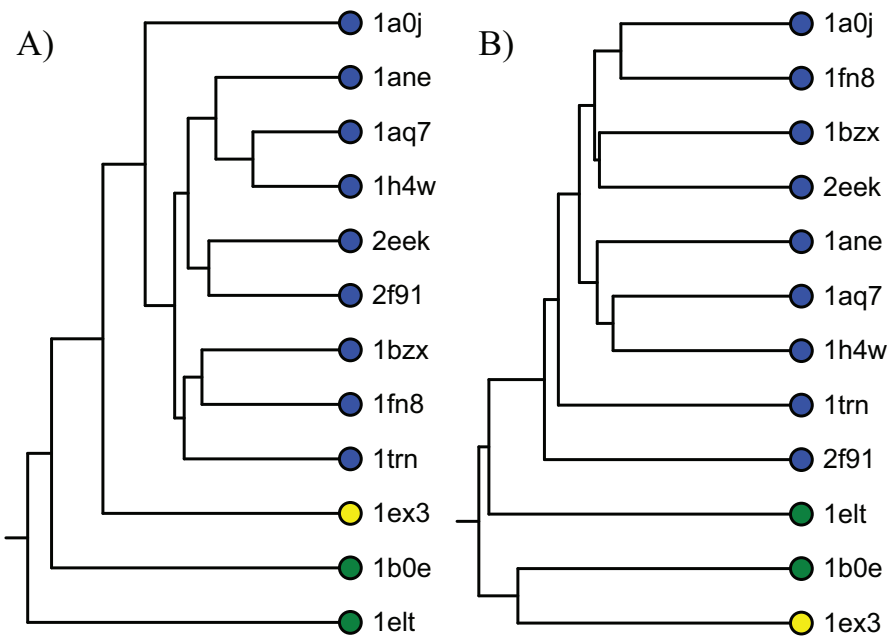

Figure 4.4: Comparison of clusterings of frequent regions and of individual cavities from serine protease structures. A) Clustering of frequent regions. B) Clustering of cavities from individual conformational samples. In both trees, topology is calculated based on volumetric distance. Coloring, which is independent of clustering topology, indicates the ligand binding preference of the protein.

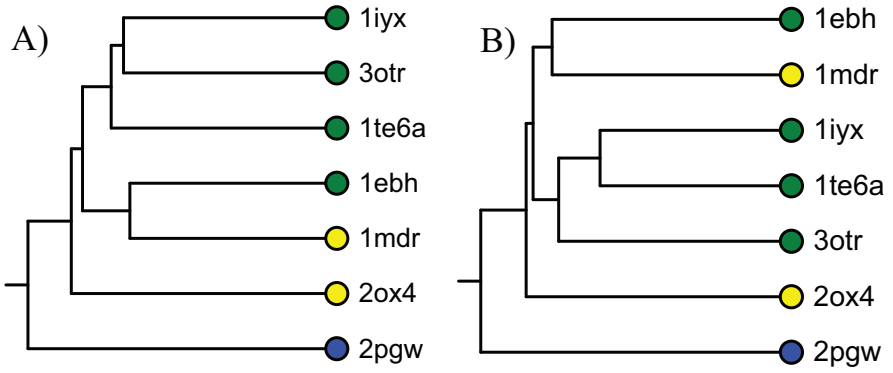

Figure 4.5: Comparison of clusterings of frequent regions and of individual cavities from enolase structures. A) Clustering of frequent regions. B) Clustering of cavities from individual conformational samples. In both trees, topology is calculated based on volumetric distance. Coloring, which is independent of clustering topology, indicates the ligand binding preference of the protein.

4.1.5 Testing FAVA

We hypothesize that similarities and differences between frequent regions can be used to classify samples of ligand binding cavities based on their binding preferences. Frequent regions of proteins with identical binding specificity are expected to be grouped into into the same cluster.

Figure 4.4A illustrates a UPGMA clustering of serine protease frequent regions based on volumetric distance. 10 out of total 12 serine protease were correctly predicted. Trypsins were correctly clustered away from other serine proteases. Elastases were also separated, but Atlantic salmon elastase was placed as an outlier because it has zero volume. Chymotrypsin was correctly separated from both trypsins and elastases. Figure 4.4B is an example of a UPGMA clustering generated from randomly selected conformational samples of each protein. We can see that one salmon elastase (pdb: 1elt) is classified as more similar to the trypsins than it is to porcine elastase (pdb: 1b0e), and that 1b0e is more similar to the chymotrypsin than anything else. This kind of miscategorization was typical of other clusteriings of cavities from randomly selected conformational samples.

A UPGMA clustering of frequent regions derived from enolase binding cavities is shown in Figure 4.5A. Totally, 6 out of total 7 enolases were correctly clustered. Frequent regions from enolase and muconate lactonizing enyzyme were correctly separated, as were frequent regions from mandelate racemase, except that the mandelate racemase from Pseudomonas putida (pdb: 1mdr) was clustered with yeast enolase instead of with mandelate racemase from Zymomonas mobilis (pdb: 2ox4). Clusterings of individual conformational samples of enolase cavities (e.g. Figure 4.5B) showed similar errors.

Overall, UPGMA clustering of frequent regions in the serine proteases and enolases generally reflected similarities and differences in specificity with equal or greater accuracy than clusterings of individual conformational samples. This result demonstrates that a flexible representation of binding cavities exhibits fewer classification errors caused by conformational flexibility.

4.1.6 Isolating Frequently Influential Amino Acids

Variations in the ligand binding cavities can cause different binding specificities and such variations can occur for many reasons. For example, the protrusion of one amino acid in one binding cavity that does not exist in another could prevent certain ligand from binding. Therefore, we are not only interested in predicting specificity within different proteins but in detecting influential amino acids that create changes of binding preferences. In this section, we explain how FAVA characterizes the flexible geometry of individual amino acids and their steric impingement on nearby binding cavities.

Given two proteins A and B, if the cavity of A is frequently different from B, then some set of amino acids is responsible for making these cavities different on a frequent basis. We identify such amino acids with FAVA.

At the level of individual samples, consider two samples of A and B, called A_i and B_j, and an amino acid r in A. We say that r makes the cavity a_i different from the cavity b_j if the intersection of the molecular surface of r in A_i, called $m(r_i)$, has a nonempty intersection with b_j. If so, then $m(r_i)$ occupies a region that is not solvent accessible in a_i but solvent accessible in b_j. Between these two samples r_i is thus one cause for the difference between a_i and b_j.

To evaluate how frequently r, an amino acid of A, creates differences between the cavities of A and B, we compute $INT_r(A, B)$, the median volume of intersection $|m(r_i) \cap b_j|$, for all pairs of samples A_i and B_j. When $INT_r(A, B)$ is large, then r frequently makes the cavity of A different from B; small values indicate that it rarely does.

4.1.7 Testing Influential Amino Acids

To evaluate how accurately FAVA can detect amino acids that create such changes, we compute the median intersection volume $INT_r(A, B)$, for all amino acids r in all elastase structures (A), and all non-elastase serine protease cavities (B). For each conformational sample of each elastase amino acid and each serine protease cavity,

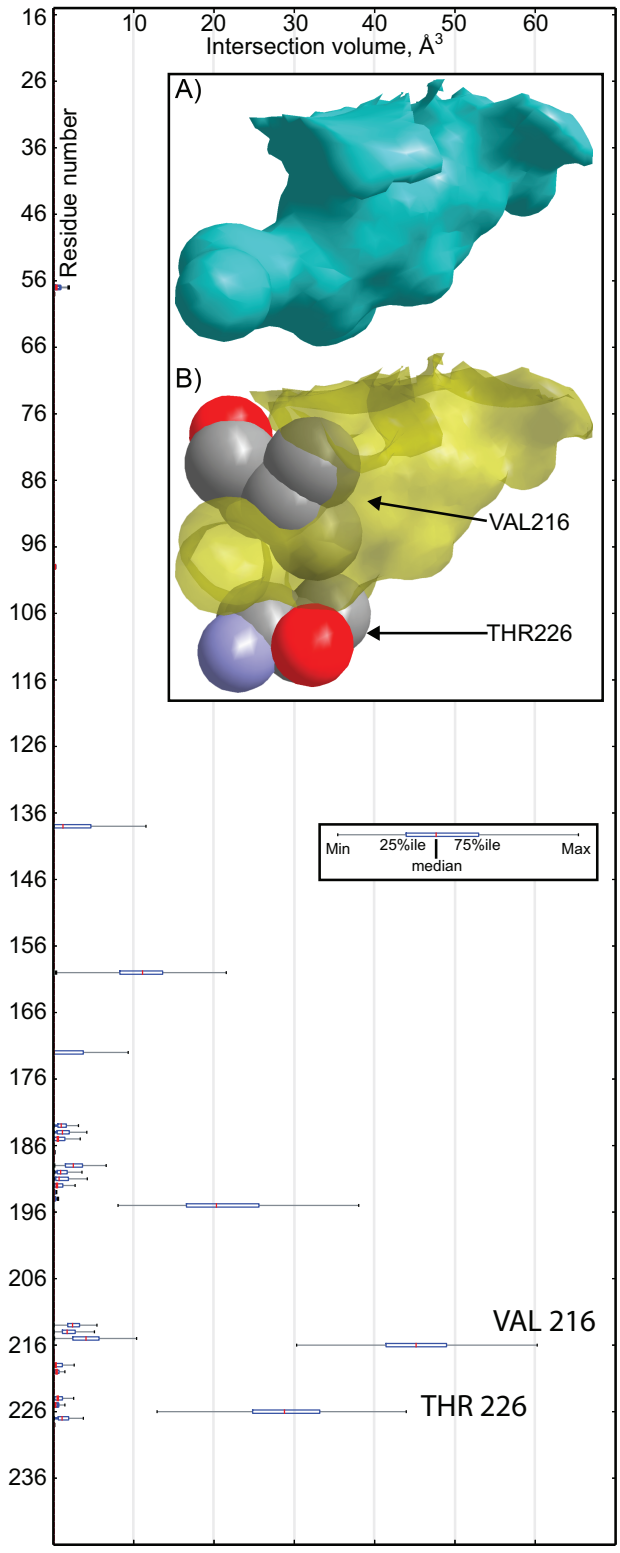

Figure 4.6: Intersection volume of amino acids from conformational samples of porcine pancreatic elastase (pdb: 1b0e) with cavities from conformational samples of salmon trypsin (pdb: 1bzx). A) The trypsin cavity (teal). B) One snapshot of Val216 and Thr226 from 1b0e, relative to the cavity.

we also measured the minimum, 25th percentile, 75th percentile, and maximum volume of intersection.

Most amino acids exhibited zero or very small intersection with any serine protease cavity, including cavities from the same protein, because the amino acid is distant from cavity. Nonetheless, some amino acids do occasionally intersect with binding cavities of the same protein. For example, among amino acids of porcine pancreatic elastase (pdb: 1b0e), the amino acid that most intersects the binding cavity of 1b0e is serine 195, the nucleophilic serine responsible for catalysis in serine proteases [132]. It occupies an median of 5 $Å^3$ inside samples of binding cavities in 1b0e.

When considering intersections between elastase amino acids and cavities from trypsins, different amino acids exhibited much larger median volumes of intersection. As an example, Figure 4.6 illustrates the degree of intersection between amino acids of porcine elastase (pdb: 1b0e) and cavities from conformational samples of salmon trypsin (pdb: 1bzx). Samples of valine 216 exhibited a median intersection volume of 45 $Å^3$ with trypsin cavities. Threonine 226 exhibited median intersection volumes of 29 $Å^3$. These findings correspond to experimental verification: Both V216 and T226 are known to occupy parts of the S1 pocket (inset, Figure 4.6), shortening it accommodate small hydrophobic amino acids [133]. We observed similar volumes of intersection between elastase amino acids and other trypsin cavities as well.

Finally, we also measured median intersection volumes between elastase amino acids and the sampled cavities of bovine chymotrypsinogen (pdb: 1ex3). Again, most amino acids exhibited small or zero median volumes of intersection with cavity samples. Serine 195, valine 216 and threonine 226 exhibited larger median volumes, at 16 $Å^3$, 20 $Å^3$, and 15 $Å^3$, respectively. These results again illustrate that amino acids that alter cavity geometry can be detected despite conformational flexibility in both the amino acids and the cavity.

4.2 PEAP: A Point-based Ensemble for Aggregate Prediction

Although FAVA proved to be a conformationally robust method for comparing protein binding cavities, it has its limits. For example, clustering results on serine proteases showed that the Atlantic salmon elastase (pdb: 1elt) was separated from another elastase protein (pdb: 1$b0e$). This error was caused because the binding cvity of 1elt is highly flexible. The flexibility makes 1elt incomparable when the frequent region does not exist and its volume becomes zero as shown in Figure 4.3. To solve this problem, we introduce a second method PEAP that employs ensemble clustering techniques to examine if it could better predict binding preferences.

4.2.1 Method Overview

To avoid the empty frequent region generated by solid representations of molecular surfaces, PEAP turns to the point-based representations as described in section . We enjoy several advantages using point-based representations. First, atomic points of amino acids selected by PEAP are all adjacent to the ligand, and they could characterize the shape of the binding site and the specificity. Second, PEAP requires selected amino acids to be k-sized, and the same length of atomic points makes it robustly comparable among all protein structures.

Overall, PEAP takes the same input as FAVA, which are conformational samples of one family of protein structures. First, we designate one protein structure as the template and explain how to compute a special substructure *structural motif*: the amino acids that are adjacent to the ligand surface. We also call it *template motif* because this is the structural motif of the template structure. The template motif is close to the binding cavity and its motion may alter the shape of the binding site. Second, we describe how we compute structural matches to identify similar substructures in other input proteins, generating the *propagated motif* for each family member.

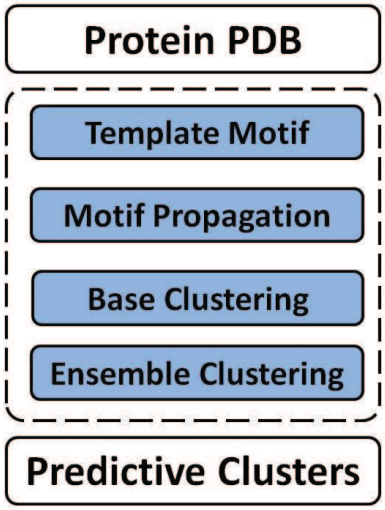

Figure 4.7: The ensemble clustering based prediction pipeline.

Given one sampled conformation of each protein, we compute all-against-all least root mean square distance (LRMSD) similarities between propagated motifs. These similarities create geometric feature vectors that correspond to high dimensional points in the geometric feature space. We continue to build a hierarchical clustering from geometric features, which outputs a clustering label one each protein.

Due to the nondeterministic nature of protein conformation sampling, the clustering could be highly unstable and no single clustering is guaranteed to be reliable across all conformational samples of all protein structures. Ensemble learning [134], as a one way to mitigate prediction errors caused by randomness, integrate multiple base methods to obtain better predictive performance than could come from any of the constituent performance alone. Therefore, we applied ensemble clustering techniques. Given a set of base clusterings, ensemble clustering output a consensus prediction that shares as much information as possible with all base clusterings [135]. Finally, we discuss how to compute such a consensus clustering to predict ligand binding specificity. The schedule is shown in Figure 4.7 and each step is detailed in section 4.2.2-4.2.4.

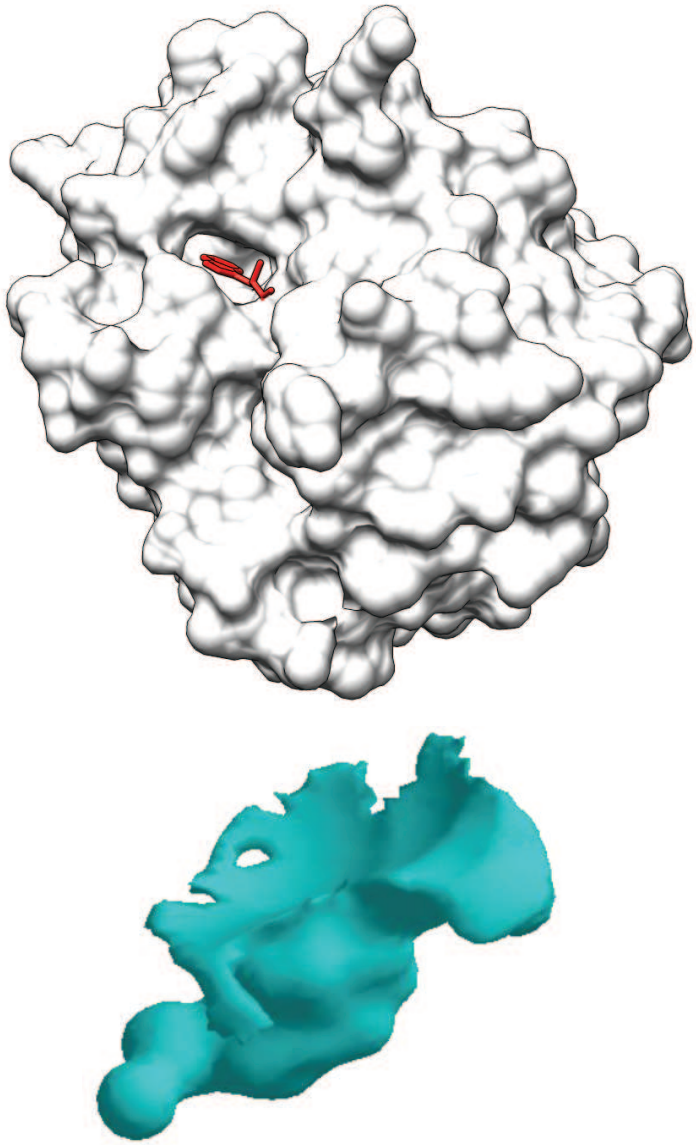

Figure 4.8: The molecular surface of a cold-adapted fish species trypsin (pdb:1a0j) with its respect to its binding ligand (red stick). The solid representation of the binding cavity generated by VASP is shown in teal region

4.2.2 Structural Motif Construction

Template Motif Construction

Within the input family of proteins, we select one structure T as the template and refer to its conformational samples as $\{T_1,T_2,...,T_n\}$. We define the molecular surface of each sampled ligand binding cavity as $\{t_1,t_2,...,t_n\}$ using VASP [43]. The binding cavity of protein 1a0j, as an example, is illustrated in Figure 4.8. Following our earlier work [128] as described in Section 4.1, we compute the average intersection volume of each amino acid r between the sampled amino acid r_i of T_i and the sampled ligand binding cavity t_j for all pairs of samples i and j. The large average intersection volume indicates that r frequently changes the shape of the binding cavity. In this work, we rank all the amino acids by their average intersection volume and return the top k as the template motif $S = \{S_1, S_2, ..., S_k\}$ where S_x is the sequence number for the xth top amino acid. The positions of motif S characterize the atomic geometry of the binding site. It is noted that our method is independent of intersection volume calculation and adapting other reasonable motif generation methods, such as expert manual selection [39, 61, 63], literature search [65] and motif database retrieval [136, 66, 67, 68], could also be successful.

Motif Propagation

The computed template motif S is matched against a family of protein structures $F = \{f_1, f_2, ..., f_m\}$, yielding a set of matches $\mathbf{M}_{S\to F} = \{M_{S\to f_1}, M_{S\to f_2}, ..., M_{S\to f_m}\}$. In this work, FATCAT [95] is used between the template structure T and each protein structure f_i to identify substructure matches by searching every amino acid in motif S and returning the matched amino acid in f_i. FATCAT was selected because of the availability and compatibility to flexible structure comparisons. Every substructure match $M_{S\to f_i}$ is a mapping between S and a substructure of f_i, and all the amino acids in the substructure are returned as a propagated motif, S_{f_i}. If any arbitrary amino acid S_i in S is aligned to a gap, S_i will be removed from the template motif.

Substructure Difference: A Case Study

We selected $1a0j$ as the template structure for serine protease superfamily and $1ebh$ as the template structure for the enolase superfamily. All the structures in the same superfamily have identical protein folds and the choice of the template structure is of little difference in generating propagated motifs. We ranked all the amino acids by the average intersection volume with the binding cavity and selected top 8 amino acids as a case study. In table 4.1, we show all amino acids in the template motif. Figure 4.9A illustrates the atomic positions of the template motif in one conformational sample structure of protein $1a0j$. It is observed that all amino acids are close to the ligand, and their motions may enlarge, shrink or even separate binding cavities. Figure 4.9B-D illustrate the motif propagation by detecting substructure matches by whole structure superposition. We hypothesize that geometric differences of structural motifs cause different specificities.

4.2.3 Base Clustering Generation

To create a base clustering, we take a random sampling of protein conformations $F' = \{f_{1_i}, f_{2_{i'}}, ..., f_{m_{i''}}\}$ as input where f_{x_y} indicates the yth conformation of the structure f_x. All these conformations are superposed onto one selected structure f_x by minimizing the overall root mean square distance (RMSD). We write the pairwise LRMSD between two propagated motifs as $L(S_{f_j}, S_{f_k})$. The LRMSD is obtained by computing C_α atom RMSD of all amino acids in propagated motifs on F'. The geometric feature $\mathbf{g}_j$ for protein f_j is a vector defined as $\mathbf{g}_j = \{L(S_{f_j}, S_{f_1}), ..., L(S_{f_j}, S_{f_m})\}$. The geometric feature space of all-against-all LRMSD alignment within a protein family can be represented by a matrix $\mathbf{G} = \{\mathbf{g}_1, ..., \mathbf{g}_m\}$. Each $\mathbf{g}_j$ is a point in the feature space. We hypothesize that proteins with identical

PDB	Motifs
1a0j	S190 G193 S195 V213 W215 G216 K224 P225
1ebh	D246 C247 Q295 D320 K345 H373 R374 K396

Table 4.1: The template motif

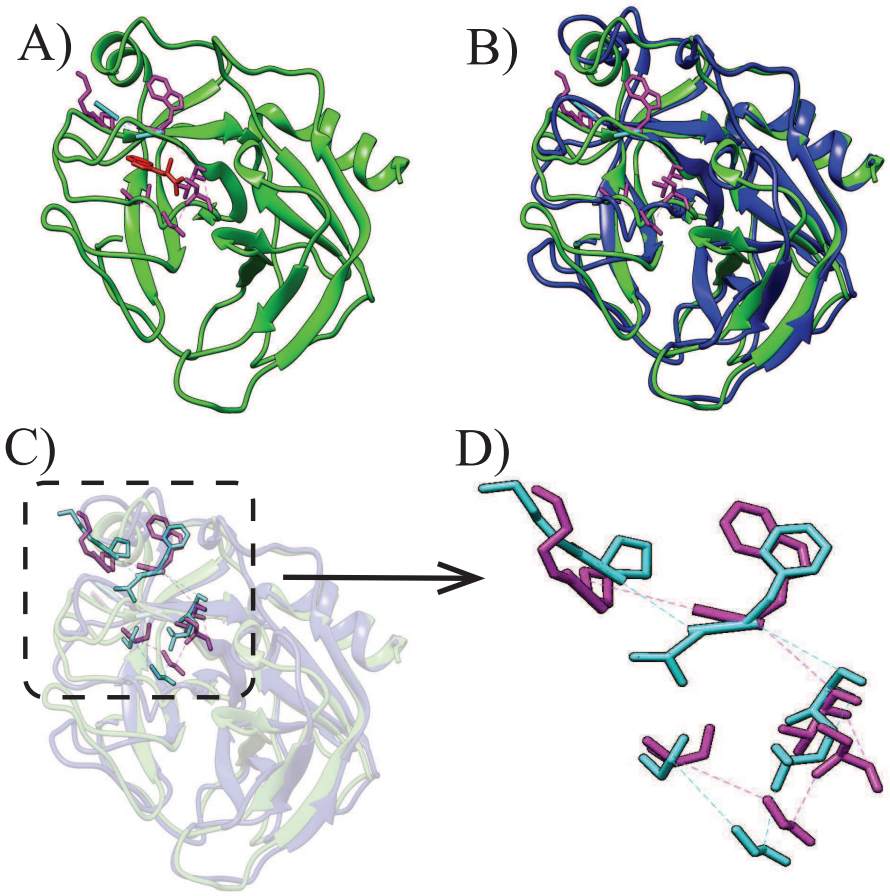

Figure 4.9: An example of the template motif in a cold-adapted fish species trypsin (pdb:1a0j) and the motif propagation to the porcine pancreatic elastase (pdb:1b0e). A) The structure of the template motif (pink sticks) in protein 1a0j (green) where the binding ligand is shown in red sticks. B) Protein 1b0e (blue) is structurally superposed onto 1a0j using FATCAT. C-D) The motif propagation by detecting amino acids (teal stick) that matches to each amino acid in the template motif.

binding specificity should be nearby in the feature space and be clustered into the same group.

To test our hypothesis, we continue to use the UPGMA to generate a base clustering using geometric features. The UPGMA outputs one base clustering as a label vector λ by specifying the number of clusters where the ith element $\lambda_i \in \{1, 2..., c\}$ indicates the cluster assignment for each feature $\mathbf{g}_i$.

4.2.4 Ensemble Clustering

In this step, we have r base clustering vectors $\{\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(r)}\}$ using conformation sampling with replacement. In order to ensemble all the base clusterings, we need a combination function Γ to create a consensus clustering $\lambda^* = \Gamma(\{\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(r)}\})$. Given m protein structures and n conformations of each structure, the number of all possible samplings is n^m, and the exponential size of combination is impractical even for very small n and m. Therefore, the brute force search over all possible samplings is infeasible and a heuristic strategy is needed.

Here, we adopt a cluster-based similarity partitioning algorithm (CSPA) to compute a consensus clustering. Essentially, if two objects are in the same cluster, they are considered to be fully similar, and if not they are fully dissimilar. To achieve this, we convert a base clustering vector $\lambda^{(q)}$ of size m to an $m \times m$ base similarity matrix $\mathbf{M}^{(q)}$ by:

$$\mathbf{M}^{(q)}_{(i,j)} = \begin{cases} 0 & \text{if } \lambda^{(q)}_i = \lambda^{(q)}_j \\ 1 & \text{otherwise} \end{cases} \tag{4.2}$$

we average all the base similarity matrices, yielding an averaged similarity matrix $\mathbf{M}$. Here, the less $\mathbf{M}_{(i,j)}$ is, the more possibility that the ith object and the jth object will be grouped into the same cluster. Finally, we form a consensus UPGMA clustering based on the averaged similarity matrix. The general process of CSPA is illustrated in Figure 4.10. For more details about ensemble clusterings and CSPA, see [137, 135].

Input: Data set $\mathbf{G} = \{\mathbf{g}_1, ..., \mathbf{g}_m\}$;
Base UPGMA clusters $\Omega^{(q)}$, $q = 1, ..., r$;
The consensus UPGMA cluster Ω;
Process:
1. **For** $q = 1, ..., r$:
2. $\lambda^{(q)} = \Omega^{(q)}(\mathbf{G})$;
3. Form an $m \times m$ base similarity matrix $\mathbf{M}^{(q)}$ from $\lambda^{(q)}$ using Equation (4.2);
4. **End**
5. $\mathbf{M} = \frac{1}{r} \sum_{q=1}^{r} \mathbf{M}^{(q)}$;
6. $\lambda^* = \Omega(\mathbf{M})$;
Output: Ensemble clustering vector λ^*.

Figure 4.10: CSPA Ensemble Clustering Algorithm.

4.2.5 Testing PEAP

Figure 4.11A illustrates superposition of propagated motifs generated by top 8 amino acids in serine proteases using selected conformational samples for each protein subfamily. The superposition exhibits geometric diversities and motif structures in proteins with the same binding specificity tend to form closely-located substructure clusters. Atomic positions of Glycine 193, as a notable example, in trypsin samples (green sticks in the dotted box) are separated from those of elastase and chymotrypsin samples, and consequently Glycine 193 could an effective marker for discriminative analysis of binding specificity. The structural motif superposition in the enolases is shown in Figure 4.11B. Only 6 amino acids were selected because Aspartate 246 and Cysteine 247 were aligned to gaps during the motif propagation and were removed from the template motif. Again, closely-located substructure clusters that indicate specificity can be detected. These observations show that motif propagation could identify subtle variations in local structures to compare proteins that have identical folds but bind to different substrates.

We expect that proteins with the same binding preference are grouped into the same cluster in PEAP prediction. We also expect that PEAP could enhance specificity prediction compared to FAVA. To test PEAP, we continue to select $1a0j$ and $1ebh$ as the template structure in each protein superfamily and choose top 8 amino

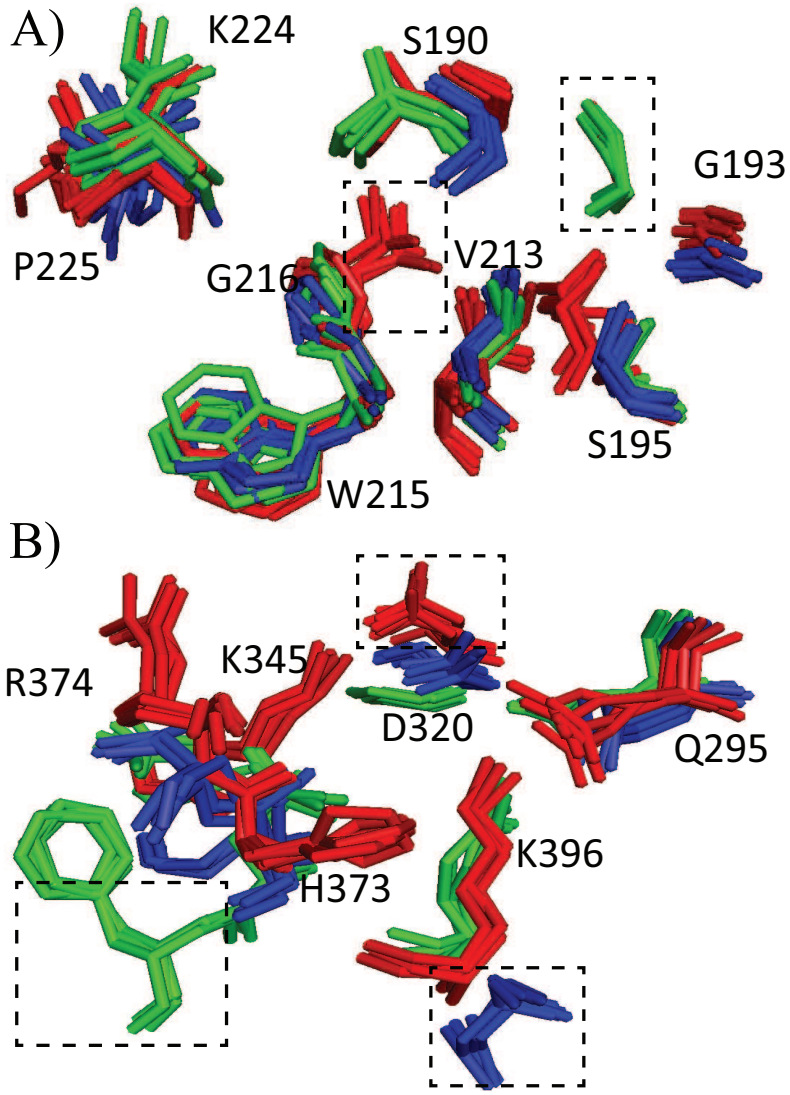

Figure 4.11: Superposition of sampled template motifs and propagated motifs of serine proteases shown in A) and the enolases shown in B) where 5 samples were randomly selected for each protein subfamily. The corlor of each aligned substructure indicate the ligand binding specificity. Substructures in propagated motifs of proteins with identical binding specificity can group into structurally co-located clusters (dotted rectangle). The figure is generated with Pymol [2].

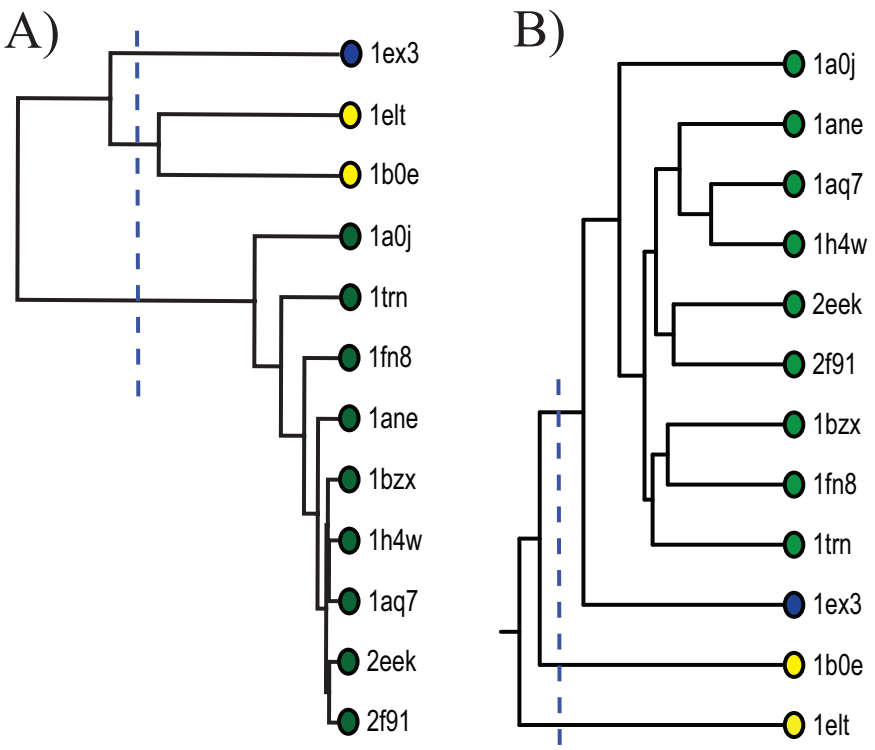

Figure 4.12: Comparison of UPGMA clustering of the ensemble method and of FAVA from serine proteases. A) Clustering of the ensemble method using propagated motifs. C) Clustering of frequent regions using FAVA. In both UPGMA trees, the dotted blue line is used to specify the number of subfamilies to generate predictive clusters. Coloring, which is independent of clustering topology, indicates the ligand binding preference of each protein.

acids with the largest average intersection volume with the binding cavity. Figure 4.12A demonstrates the ensemble UPGMA clustering of propagated motifs on serine protease structures. Proteins in the same subfamily are correctly clustered into the same group. Figure 4.12B demonstrates the UPGMA clustering of frequent regions using FAVA. We observe that the only chymotrypsin protein, $1ex3$, was misclassified into the trypsin cluster and two elastases were separated into different clusters. Moreover, ensemble UPGMA clustering exhibits a greater similarity between trypsins than FAVA clustering. This indicates that structural motifs could be better markers to distinguish proteins with different binding preferences.

Figure 4.13A shows the ensemble clustering of propagated motifs on enolase superfamily structures. Three subfamilies are all correctly clustered by their binding

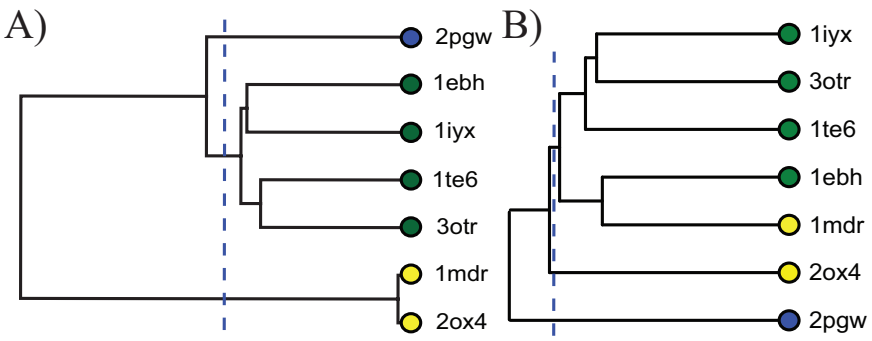

Figure 4.13: Comparison of UPGMA clustering of the ensemble method and of FAVA from the enolases. A) Clustering of the ensemble method using propagated motifs. C) Clustering of frequent regions using FAVA. In both UPGMA trees, the dotted blue line is used to specify the number of subfamilies to generate predictive clusters. Coloring, which is independent of clustering topology, indicates the ligand binding preference of each protein.

specificities. Figure 4.13B shows the clustering using FAVA. We can see that two mandelate racemases were separated and one of them, $1mdr$, was misclassified into the enolase subfamily cluster. Similarly, greater similarities between the enolases and between mandelate racemases was detected using the ensemble clustering.

Overall, UPGMA clusterings reveal that PEAP improves the specificity prediction and it could be a robust tool for flexible protein structure comparisons despite great conformational flexibilities in the binding cavity.

4.3 Conclusion

In this chapter, we have presented two novel approaches to solve the problem of aggregate prediction that have been defined in section 1.3.2. FAVA is the first conformationally general method to compare proteins with identical folds but different binding specificities. FAVA permits detailed volumetric comparisons of binding cavities despite considerable structural variations. PEAP presents a computational tool to compare protein cavities based on ensemble clustering. Different from FAVA with solid representation of molecular surfaces, PEAP identifies functional atomic points to avoid empty frequent regions. PEAP also ensembles multiple base clusterings for a consensus clustering to predict the binding specificity.

We evaluated FAVA and PEAP on serine protease and enolase superfamilies. FAVA was able to classify members of both superfamilies with equal or superior performance than classifications where only a single conformation had to be selected at random. Measuring the median volume of intersection between sampled amino acids of one protein and the sampled cavities of another, FAVA was capable of identifying amino acids that have an experimentally established influence on binding specificity. Experimental results of PEAP revealed that all protein structures in both superfamilies can be correctly predicted. The results indicate that PEAP enhances the specificity prediction by mitigating prediction errors.

As practical tools for flexible comparisons of the binding cavity, both FAVA and PEAP have considerable potentials for wider applications. In many cases, efforts to create proteins with engineered binding preferences already involve the simulation of protein structures. Our methods introduce an analysis of the resulting simulation data that might yield more detailed comparisons of frequently conserved regions or selected amino acids, which can be changed for a desired binding preference.

Chapter 5

Individual Prediction Pipelines Development

In this chapter, we introduce methods that focus on the individual prediction problem. Individual prediction is important because an analysis of individual snapshots might, for example, reveal subtle structural dissimilarities between proteins with different specificities, and these st ructural variations could be further altered to change specificity of each conformation. The individual prediction takes each conformation as an independent source while, in aggregate prediction, all conformations of the same protein are taken as one unit of input. The expected output that solves the individual prediction problem is grouping each protein snapshot into categories with different specificities.

Here, we introduce three representations, an atomic point representation, a volumetric lattice representation and an electrostatic lattice representation, for representing protein flexibility. The atomic point representation identifies positions of selected amino acids to describe geometric motions in the binding site. The volumetric lattice representation calculates cavity volume in a user-defined cube within the binding site. The electrostatic lattice representation computes the volume of protein electrostatic isopotentials in a a user-defined cube. All these representations convert each protein conformation into a feature vector.

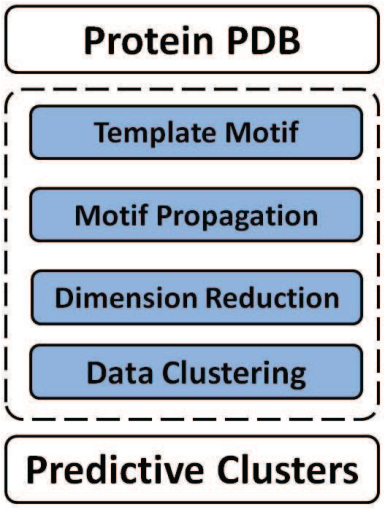

Figure 5.1: The atomic point representation pipeline.

As we have discussed in section 1.3.2, small motions that come from a 100 nanosecond simulation used in this thesis do not change specificity and each protein conformation thus takes the same specificity as the native structure. To test all three representative models, we hypothesize that conformations of proteins with identical binding specificity are close together while conformations with different specificities separate apart. We evaluate these methods on the same data sets, serine proteases and the enolases.

5.1 An Atomic Point Representation

5.1.1 Method Overview

Overall, this representation accepts conformational samples of one superfamily of protein structures as input. Each conformation of the binding cavity will be mapped into an embedded feature space and we call this the *conformation space map*. This space map enables comparisons of every protein snapshot to predict its binding specificity. First, we leverage motif propagation to detect template motifs and propagated motifs among all input protein structures as we have described in section 4.2. We extract three dimensional Carbon alpha coordinates of each amino acid in structural motifs. Therefore, each conformation can be characterized with a feature vector.

Since structural motifs are propagated by detecting substructural matches, causing some selected amino acids to be highly similar in geometry. These similar structures will increase the dimensionality of the feature vector but do not provide discriminative information on specificity. Hence, in the next step, we perform dimension reduction techniques to reduce dimensionality. Here, we select two effective reduction methods, Non-negative Matrix Factorization (NMF) and Principal Component Analysis (PCA). Finally, we perform data clustering to generate predictions that correspond to the problem of individual prediction. The outline of our approach is shown in Figure 5.1.

5.1.2 Motif Propagation

The motif propagation consists of two parts: the template motif construction and motif propagation to other protein structures. This step has been detailed in Section 4.2.2. The meaning of motif propagation is to identify selected amino acids, which are comparable among all protein conformations, for characterizing flexibilities in the binding site. We applied FATCAT [95] to identify substructure matches for motif propagation and other substructure matching algorithms, such as LabelHash [138] and Match Augmentation [63], could substitute for FATCAT as well.

5.1.3 Dimension Reduction

Once propagated motifs have been generated, one conformation of the binding cavity can be characterized as a geometric feature vector where each value is x or y or z direction coordinate of Carbon alpha atom. The feature matrix $X = \{x_1, ..., x_n\} \in \mathbb{R}^{m \times n}$ represents geometric features of all protein conformational samples where n is the total conformation number and m is the feature dimensionality, and the matrix will be taken as input for dimension reduction.

Non-negative matrix factorization (NMF) [139] is a matrix decomposition algorithm for parts-based data representation of matrices with non-negative elements. Given an input matrix X, NMF aims to find two non-negative components $W \in$

$\mathbb{R}^{m\times r}$ and $H \in \mathbb{R}^{r\times n}$ to minimize the objective function where r is the reduced dimensionality:

$$\begin{aligned} &\min_{W,H} F = ||X - WH||^2 \\ s.t. \quad &W_{ij} \geq 0, H_{ij} \geq 0 \end{aligned} \tag{5.1}$$

The objective is convex with respect to either W or H, but not convex to both together so that the global optimal is difficult to find. Starting from random initialization of W and H, Lee and Seung [140] presented an algorithm to find a local minimum by iteratively update W and H:

$$\begin{aligned} W_{ij} &= W_{ij}\frac{(XH^T)_{ij}}{(WHH^T)_{ij}} \\ H_{ij} &= H_{ij}\frac{(W^TX)_{ij}}{(W^TWH)_{ij}} \end{aligned} \tag{5.2}$$

Usually, we have $r \ll m$ and $r \ll n$. NMF can be understood as trying to discover latent structures using a few dimensions in a compressed representation. If there exists negative elements in the matrix, we add negative minimum value of the matrix to guarantee the non-negative constraint.

Principal Component Analysis (PCA) [141] is one of the most popular dimension reduction methods. PCA orthogonally project a set of data points onto a lower dimensional subspace such that variances between projected data are maximized. The projection vectors can be computed as a set of eigenvectors with top r largest eigenvalues.

Both NMF and PCA provide methods for reducing the dimensionality of feature space. We then perform clustering on both reduced latent space.

5.1.4 Cluster analysis

We apply the canonical K-means clustering to detect data clusters in the feature space. The number of clusters is equal to the number of different protein subfamilies.

K-means outputs predicted cluster label l_i on each conformational sample. The clustering is evaluated with two metrics, clustering accuracy (AC) and normalized mutual information ($\overline{MI}$).

Given the predicted cluster label l_i and the ground truth g_i defined by EC number, AC is defined as:

$$AC = \frac{\sum_{i=1}^{n} \delta(g_i, map(l_i))}{n} \tag{5.3}$$

Where $\delta(\cdot)$ is the delta function that equals to one for identical comparison and equals to zero otherwise and $map(\cdot)$ is a permutation mapping function that matches clustering label set to the equivalent label from ground truth label set as much as possible. This can be done using the Kuhn-Munkres method ([142]).

Given the set of predicted clusters C and the set of ground truth clusters C', the mutual information $MI(C, C')$ is defined as:

$$MI(C, C') = \sum_{c_i \in C, c_j \in C'} p(c_i, c_j) \cdot log_2 \frac{p(c_i, c_j)}{p(c_i) \cdot p(c_j)} \tag{5.4}$$

where $p(c_i)$ and $p(c_j)$ are the probabilities that a sampled data belongs to cluster c_i and c_j respectively and $p(c_i, c_j)$ is the joint probability that a sampled data belongs to cluster c_i and c_j at the same time. In our experiments, we use the normalized mutual information $\overline{MI}$ to scale MI between 0 and 1 as follows:

$$\overline{MI}(C, C') = \frac{MI(C, C')}{max(H(C), H(C'))} \tag{5.5}$$

where $H(C)$ and $H(C')$ are the entropies of cluster set C and C' respectively, and the entropy is defined as $H(C) = -\sum_{c_i \in C} p(c_i) log_2 p(c_i)$. $\overline{MI}$ equals to one only if two cluster sets are identical and equals to zero only if two sets are independent.

5.1.5 Comparisons with State-of-the-art Methodologies

To our best knowledge, we did not find any previous work that takes the same input and output as we do here to predict the binding specificity in an unsupervised way. Therefore, direct comparisons to existing methods are not possible here. However, to demonstrate the effectiveness of the atomic point representation, we compare with existing point based methods, which are documented structural motifs in previous literature studies.

The *catalytic triad* is a group of three amino acids that function together within the active site of serine proteases. Catalytic triads are representative examples of local substructures that indicate functionally convergent evolution. For example, subtilisin and chymotrypsin have no sequence identity and very different folds, but they exhibit the same Ser-His-Asp catalytic triad, so they bind the same enzyme inhibitor [143]. In chymotrypsin, the catalytic triad is made of histidine 57, aspartate 102 and serine 195 (Figure 5.2). The side chain of the serine 195 is bounded to the imidazole ring of the histidine which accepts a proton from serine to form a strong alkoxide nucleophile when a ligand is binding. The aspartate 102is attracted by the histidine via hydrogen bound and electrostatic interaction to make it a better proton acceptor (Figure 5.2A).

Five core amino acids directly mediate the conserved reaction within the enolase superfamily. These amino acids are Lys 164, Asp 195, Glu 221, Glu 247 and His 297 in a mandelate racemase structure (pdb:2mnr) [144], and they were used by two substructure matching algorithms, SPASM [145] and LabelHash [138], to identify members of the whole enolase superfamily with high sensitivity and specificity. We call these five amino acids the *catalytic pentad*.

We selected {57,102,195} from our only chymotrypsin structure $1ex3$ as the template motif and applied motif propagation to obtain amino acid triads in other trypsins and elastases. In the enolase superfamily, we selected another mandelate racemase structure (pdb:1mdr) as the template structure because it has identical amino acid sequence as $2mnr$. Then, we extracted the catalytic pentad and prop-

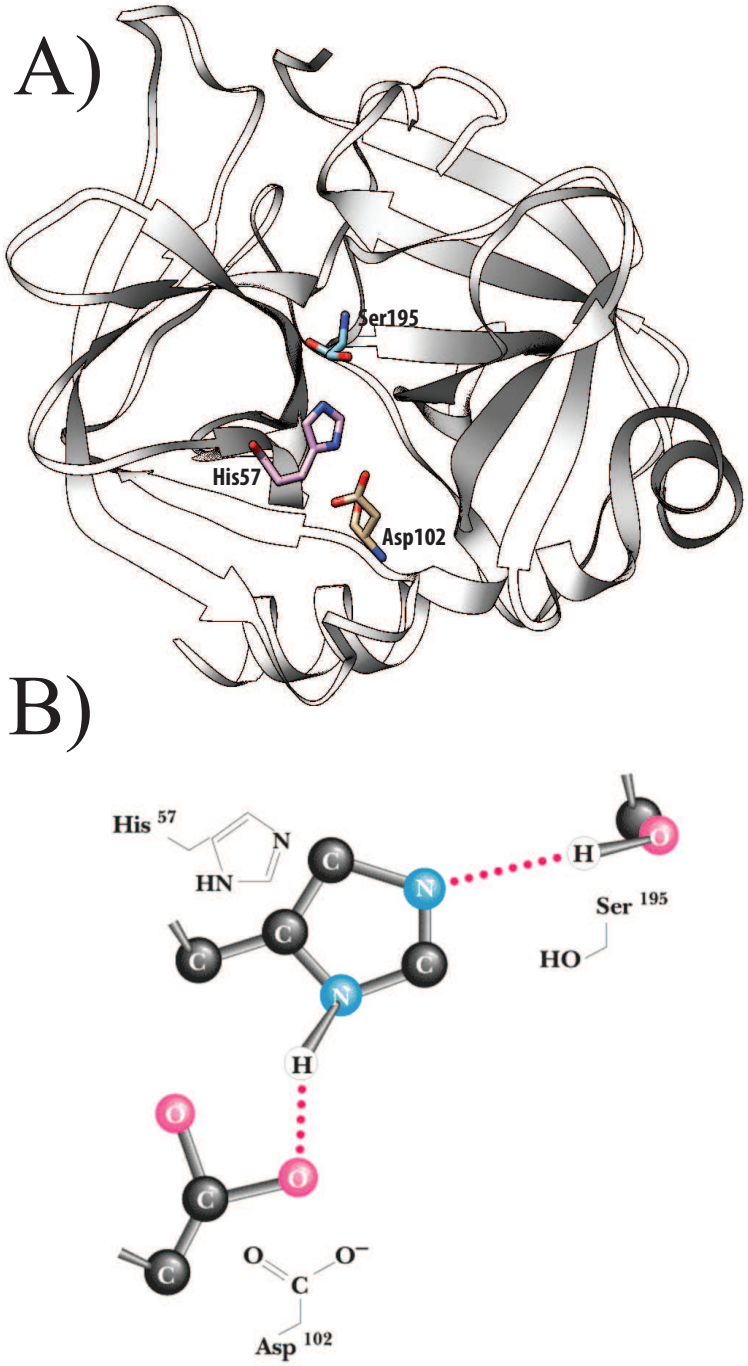

Figure 5.2: A) The Ser-His-Asp catalytic triad bound to the structure of chymotrypsins. B) The hydrogen bounds within the catalytic triad, which is illustrated in [3].

agated to other structures in the enolases. We took the catalytic triad and the catalytic pentad as baseline point based representations by selecting coordinates of Carbon Alpha atoms and compared with the atomic point representation.

5.1.6 Testing Atomic Point Representation

First, we compare the atomic point representation with baseline point based methods. Second, we visualize the conformation space map of binding cavity and evaluate its clustering performance quantitatively.

Comparing with State-of-the-art Methodologies

We selected $1a0j$ and $1ebh$ as the template structure in each protein superfamily and propagate top k amino acids which have the largest average intersection volume with the binding cavity to other proteins. Figure 5.3 and Figure 5.4 report the clustering on top k propagated amino acids with respect to the catalytic triad of serine proteases and the catalytic pentad of the enolases respectively. The number of selected amino acids k ranging between 1 and 20 was used because amino acids beyond reveal almost zero average intersection volume with the binding cavity. For each k, 20 K-means runs were conducted and the average performance was reported.

These two figures reveal several insights. First, using the atomic point representation, 98.5% of serine proteases ($k = 10$) and 85.7% of the enolases ($k = 7-11$) were correctly predicted as best results. This shows that our method is able to correctly categorize protein conformations that correspond to different binding specificities. Second, performance increased as more amino acids were added but suddenly decreased when the size of motif was larger than a threshold. This means that, if the size of the structural motif is too small, the binding cavity will be under-represented because some other influential amino acids are not included. If the motif size is too large, the binding cavity will be over-represented with systematic noises because amino acids that are irrelevant to binding are included. Third, in most k values, the motif used in our method largely overperformed the catalytic triad or the catalytic

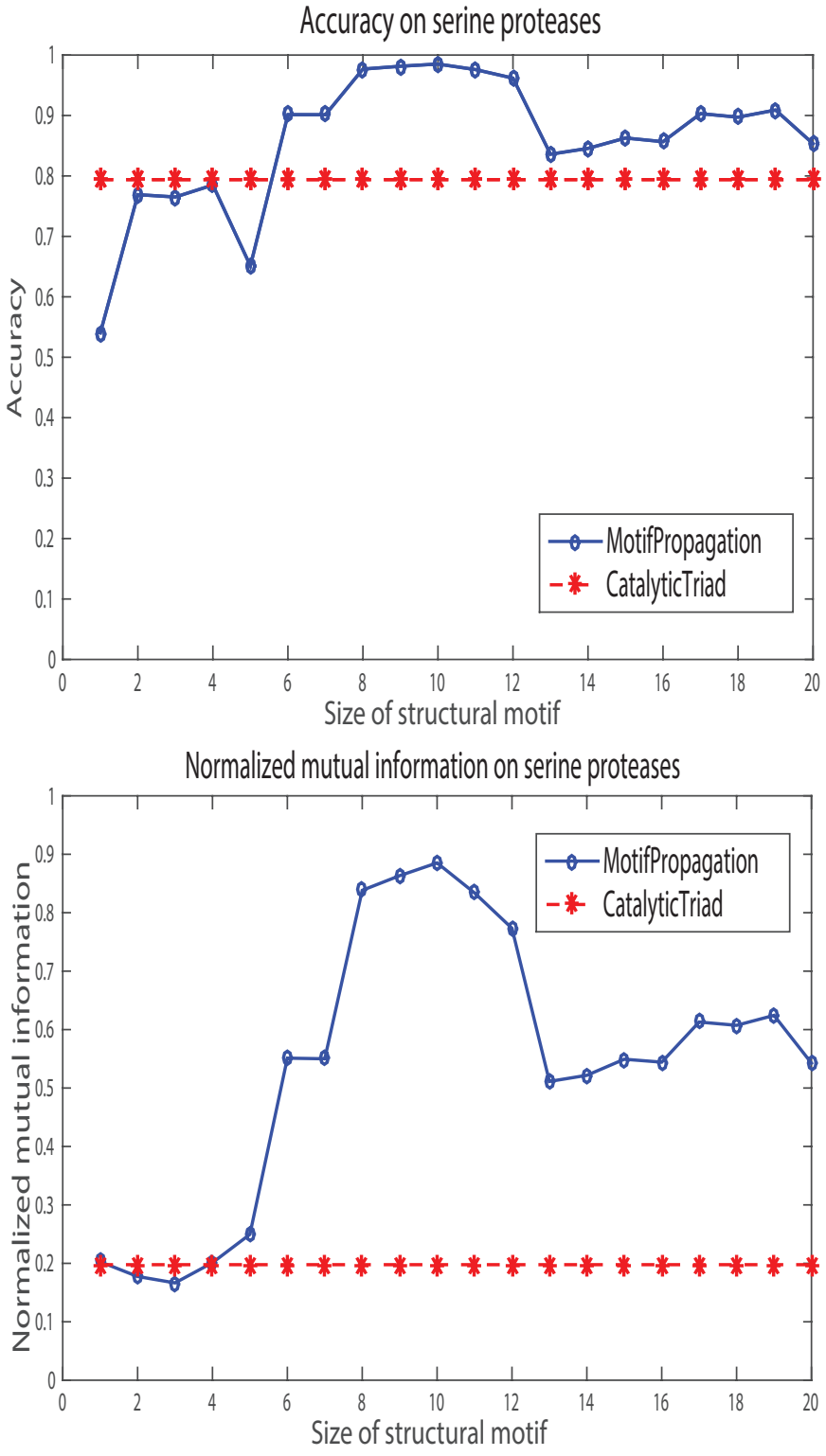

Figure 5.3: Clustering performance comparisons with the catalytic triad on serine protease superfamily.

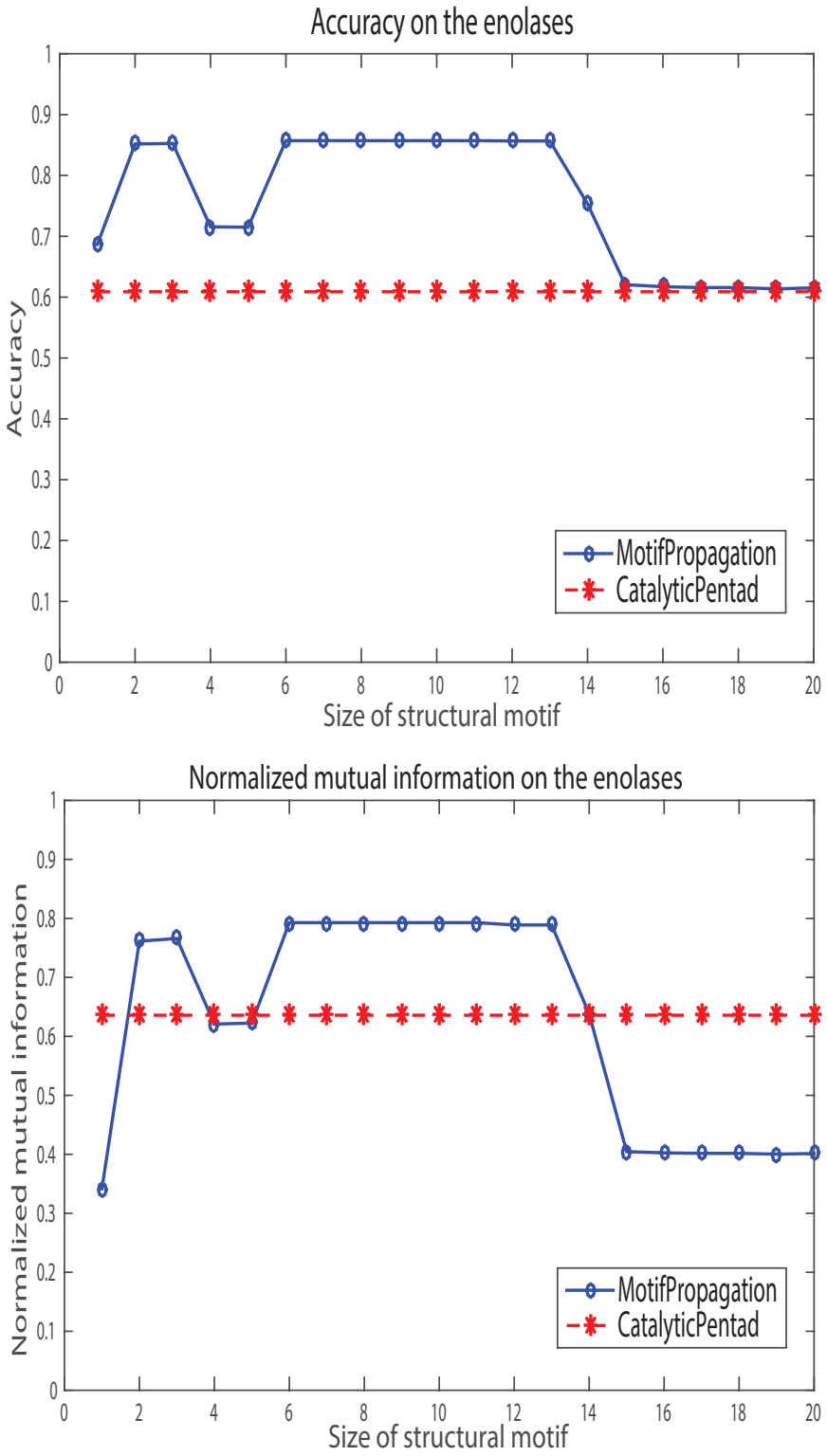

Figure 5.4: Clustering performance comparisons with the catalytic pentad on the enolase superfamily.

pentad. These documented amino acids proved to be effective in characterizing the functional site catalysis, but they are not necessarily to be the best choice for predicting specificity in the context of binding cavity motions.

Conformation Space Map Visualization

As a case study, we project each data input into a $3D$ lower space in both NMF and PCA. The conformation space map of the binding cavity on serine proteases is illustrated in Figure 5.5. It is observed that, in both NMF and PCA conformation space maps, conformational samples of proteins with identical binding preference are represented by spatially adjacent points and tend to be grouped into the same predictive cluster. These specificity-sensitive clusters can be further evaluated by comparing to ground truth values. The conformation space map on the enolases is illustrated in Figure 5.6 where a similar data distribution pattern is observed. These visualizations reveal a high level organization for classification of the binding cavity, and they provide a more straightforward system for protein classification than traditional hierarchical classifications, including EC [146], CATH [147] and SCOP [148], because our mappings are presented in a continuous space.

Evaluation in Different Feature Space

We continued to conduct clustering evaluations in three different feature space with k values. Figure 5.7 and Figure 5.8 report clustering results in the original feature space (K-means), $3D$ PCA reduced space (PCA+K-means) and $3D$ NMF reduced space (NMF+K-means) on our date sets. Since NMF is highly dependent of data initialization, 100 NMF runs with random initialization were conducted and the best K-means result was reported.

In most cases (except when k ranged between 8 to 12 on serine proteases and equalled to 13 on the enolases), PCA+K-means achieves comparable or even better performance to K-means in the original space. This suggests that PCA extracts most data variances that are sufficient enough to distinguish protein conformations

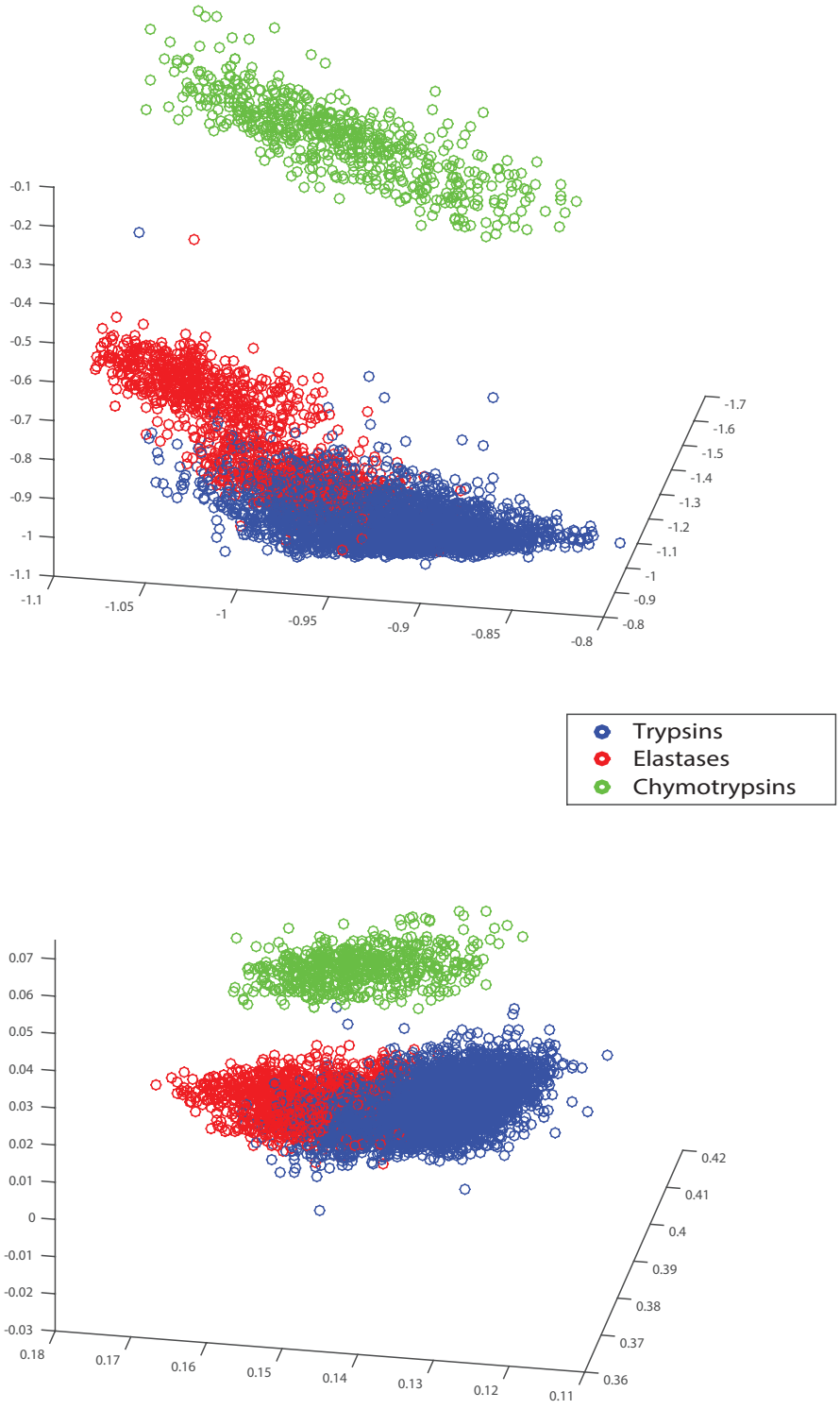

Figure 5.5: A binding cavity conformation space map of serine protease superfamily where the size of motif is set to be 8 and each protein is presented with 600 conformations. The top figure shows the NMF reduced space and the bottom figure shows the PCA reduced space. The coloring indicates the binding specificity of each conformation that is defined by EC number.

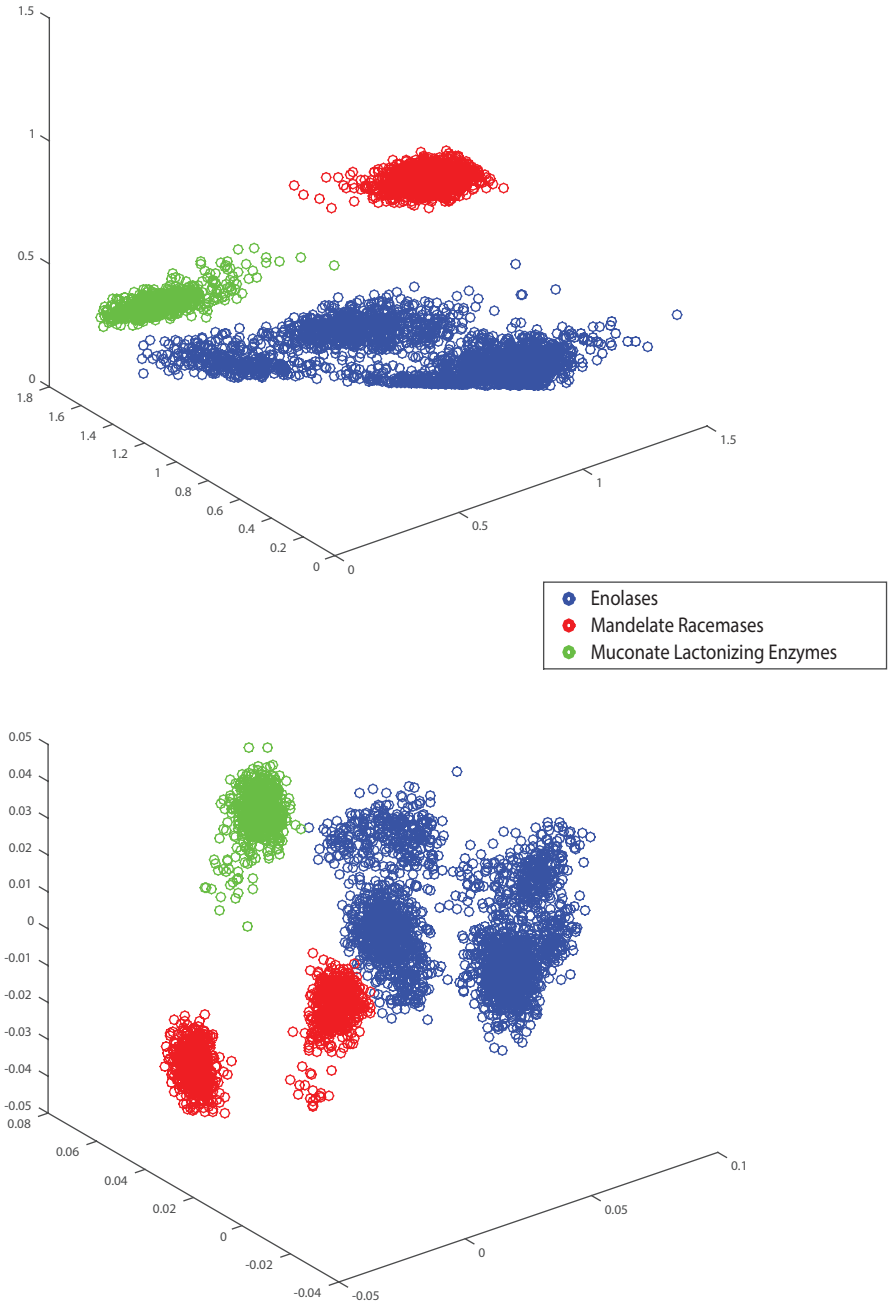

Figure 5.6: A binding cavity conformation space map of the enolase superfamily where the size of motif is set to be 8 and each protein is presented with 600 conformations. The top figure shows the NMF reduced space and the bottom figure shows the PCA reduced space. The coloring indicates the binding specificity of each conformation that is defined by EC number.

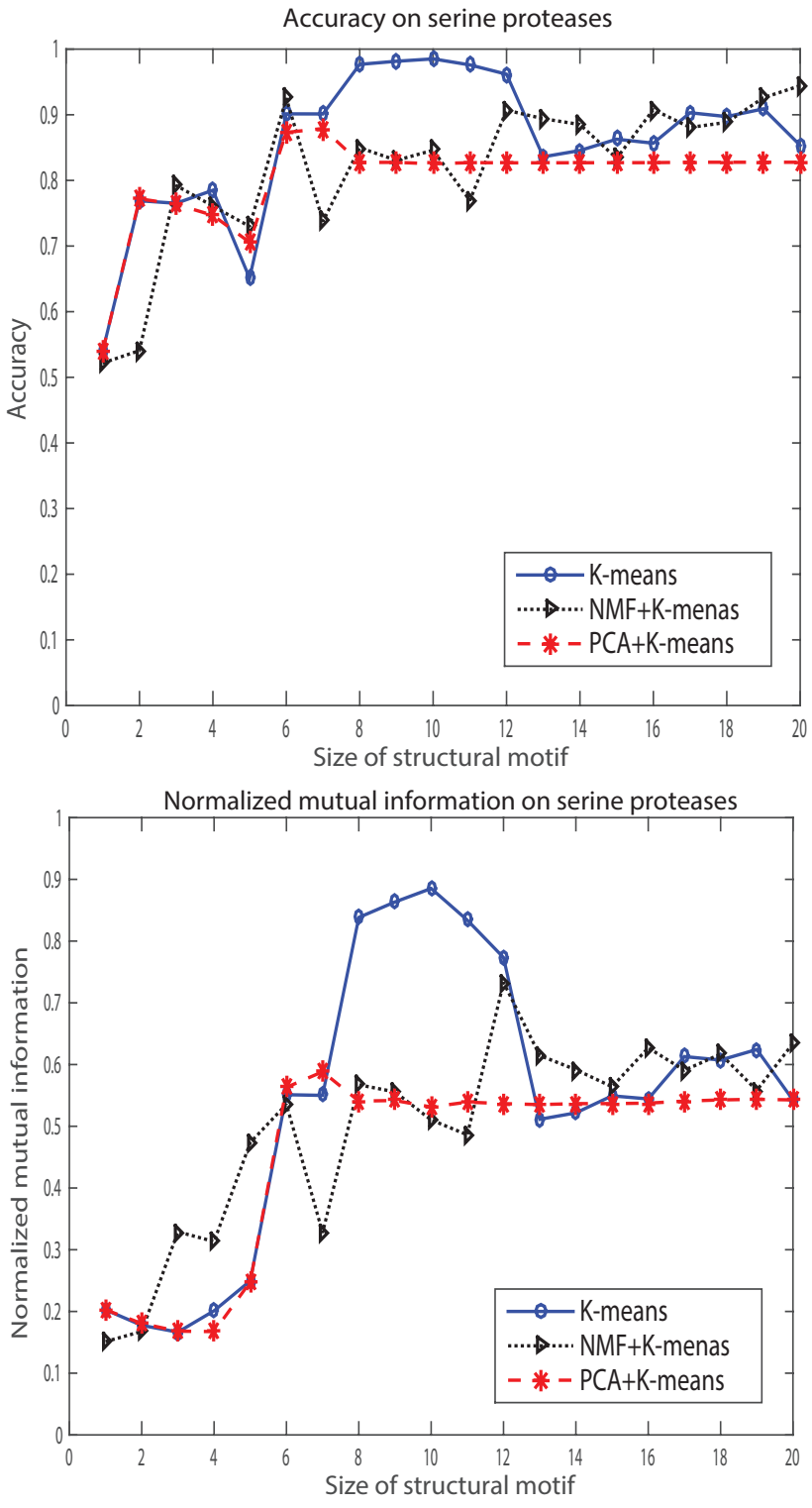

Figure 5.7: Clustering performance in three different feature space with respect to the size of structure motif on serine protease superfamily.

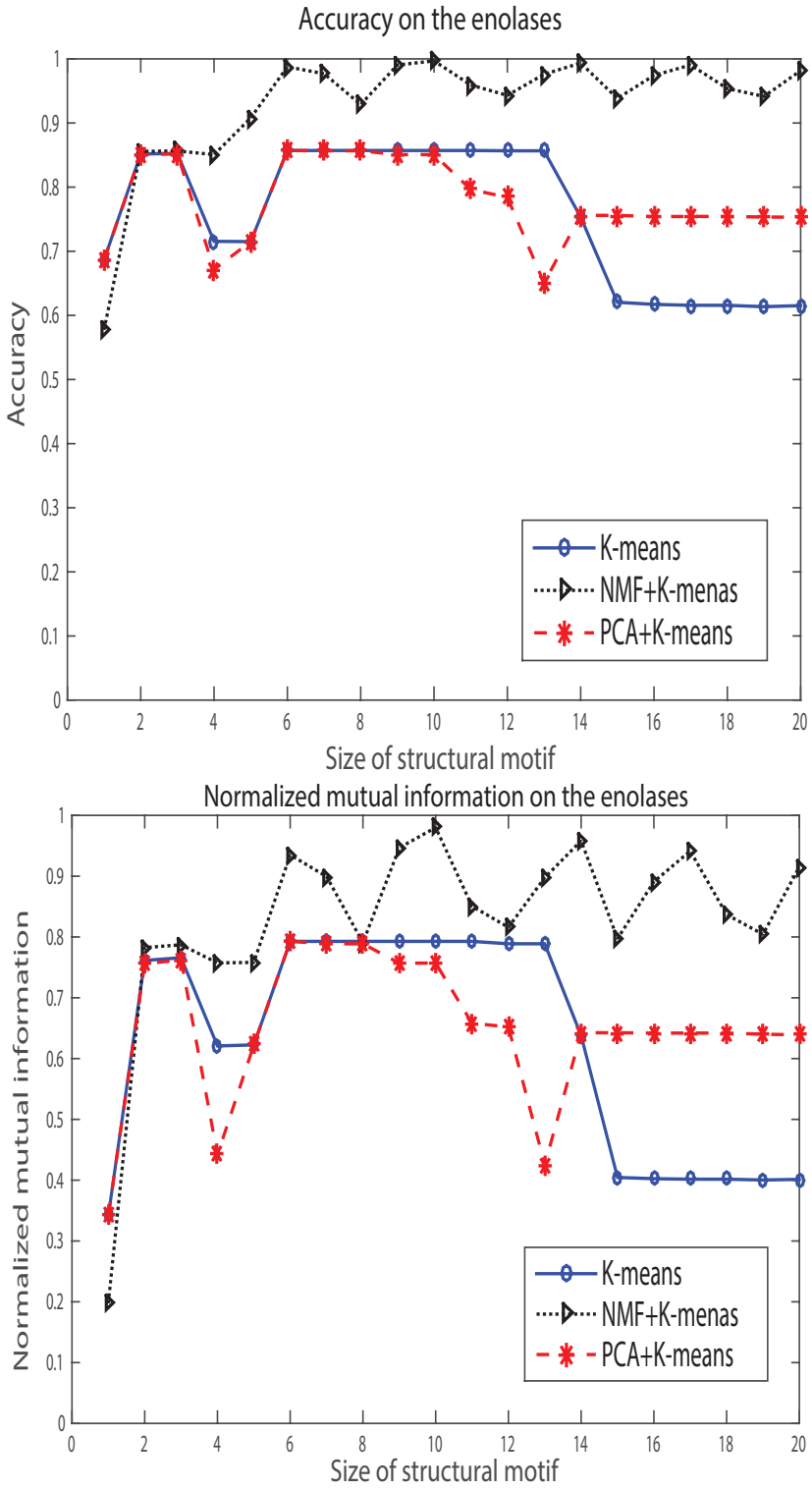

Figure 5.8: Clustering performance in three different feature space with respect to the size of structure motif on the enolase superfamily

with different specificities. NMF+K-means achieves similar performance on serine proteases except when k ranged between 7 to 12 to compare with K-means. NMF+K-means achieves obviously better performance than other two methods on the enolases. These results validate the power of NMF in identifying latent topic structures that may be embedded in the original feature space, which has been discussed in other data clustering tasks [149, 150, 151, 152].

Overall, our conformation space map reveals a high-level representation of binding cavity conformations. The clustering results show that our method is able to correctly distinguish similar proteins but with different specificities. Our method proves to be an effective tool for flexible protein structure comparisons.

5.2 A Volumetric Lattice Representation

We have presented a atomic point representation for representing the flexibility in the binding site where atomic positions of influential amino acids are extracted. However, this method only considers the coordinates of Carbon alpha atoms and ignores sidechain atoms motions around the binding cavity because sidechains of different amino acids have different number of atoms, making it impossible to find a one-to-one alignment between all atoms. Second, the atomic point representation aligns protein substructures by positions of atoms rather than comparisons of the shape of the binding cavity. However, it is the open space within the cavity that accommodates binding partners and the similarity of that space provides more direct evidence for the same binding specificity. To deal with this issue, this section will introduce a volumetric lattice representation which aims for an all-atom representation for the binding cavity.

5.2.1 Method Overview

First, for every conformational sample, we define the shape of the ligand binding cavity as a solid object. The solid representation is selected because it describes geometries of all adjacent atoms that could sterically hinder binding while the atomic

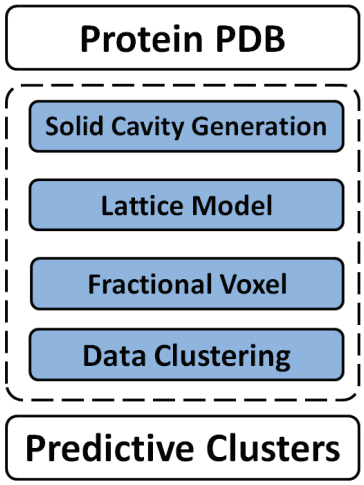

Figure 5.9: The volumetric lattice representation pipeline.

point representation only considers Carbon alpha atoms. Second, we describe how we build a lattice model on binding cavity solids with many user-defined cubes. The *volumetric voxel*, which is the cavity volume in each cube, is then calculated in order to localize geometric changes in each spatial unit. Volumetric voxels resemble digital image pixels that describe color values in every physical point. All volumetric voxels can be combined into a feature vector and each binding cavity conformation is then represented as a high dimensional point. In the last step, we perform K-means clustering to predict specificity on each conformation. The overview of the volumetric lattice representation is shown in Figure 5.9

5.2.2 Solid Binding Cavity Generation

Given the PDB structure of each protein conformation and a binding ligand, we compute the space of the binding cavity as a solid in the form of triangle meshes. We have already described this part step by step in Section 4.1.2. The solid representation enables a direct comparison of the shape of the binding cavity. The binding cavity of protein $1a0j$ has been illustrated in Figure 4.8.

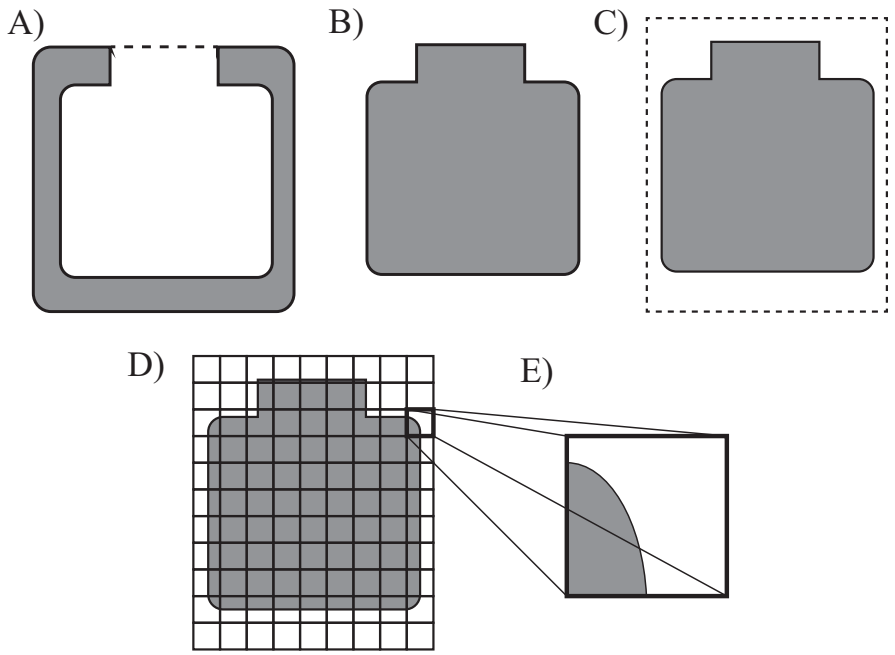

Figure 5.10: The lattice model construction. A) The CSG operations used by VASP, with input regions (light grey, dotted outline) and output regions (solid outline). B) The molecular surface of a given conformation sample(grey region) with respect to the binding border (dotted line). C) The solid representation of the binding site. D) The bounding cuboid that covers the binding cavity. E) The cubic lattice inside the bounding cuboid. F) Volume calculation in each lattice cube.

5.2.3 The Lattice Model Construction

As input, we need solid representations of binding cavities of all protein conformations(Figure 5.10B) and a lattice resolution r. First, a bounding cuboid (Figure 5.10C) is constructed to cover all binding cavity solids where the length, the width and the height are all integral multiples of r. Second, we build an axis aligned cubic lattice (Figure 5.10D) inside the bounding box so that each cube has the identical size length of r. The lattice can interpreted as a grid of lattice points that are equally spaced along the primary axes, or as a set of lattice segments that connect the co-axial lattice points, or as a collection of lattice cubes that adjacent cubes share four lattice segments. Last, using the Surveyor's formula ([153]), we measure the cavity volume of every snapshot in each lattice cube (Figure 5.10E).

5.2.4 Cluster Analysis

Given volumetric voxel representation, each conformation of the binding cavity can be characterized as a geometric feature vector $x_i \in \mathbb{R}^m$ where the feature value computes the cavity volume in one cube and m is the total number of cubes within the bounding box. All feature vectors are normalized so that each data point has unit norm. The feature matrix $X = \{x_1, ..., x_n\}$ thus represents all protein conformations, and it will be taken as input for data clustering.

We continue to perform the canonical K-means clustering, which is evaluated using clustering accuracy (AC) and normalized mutual information ($\overline{MI}$) as described before.

5.2.5 Testing Volumetric Lattice representation

We compare the volumetric lattice representation with the atomic point representation. In the atomic point representation, we select top 20 amino acids with largest average intersection volume with the binding cavity (The pseudomonas mandelate racemase, as an example, is illustrated in Figure 3.3A). In order to identify geometric features among all possible combinations of amino acids, all k-sized combinations (i.e., $\binom{20}{k}$ combinations) are generated on all possible selections of k. For example, when $k = 10$, all $\binom{20}{10} = 184756$ feature subsets are generated. Then, clustering performance on each subset is examined. The strategy of k-sized combinations is used because all possible geometric features get considered and the volumetric lattice representation is not compared by chance.

For each feature subset, 20 K-means runs with random initialization were conducted and their average performance was reported. Figure 5.11 and Figure 5.12 illustrate clustering performance of the volumetric lattice representation where the lattice resolution equals to 0.50. Given a variety of k values, the performance of the atomic point representation is shown in boxplot on all k-sized subsets while the performance of the volumetric lattice representation is constant because it is independent of the number of amino acids.

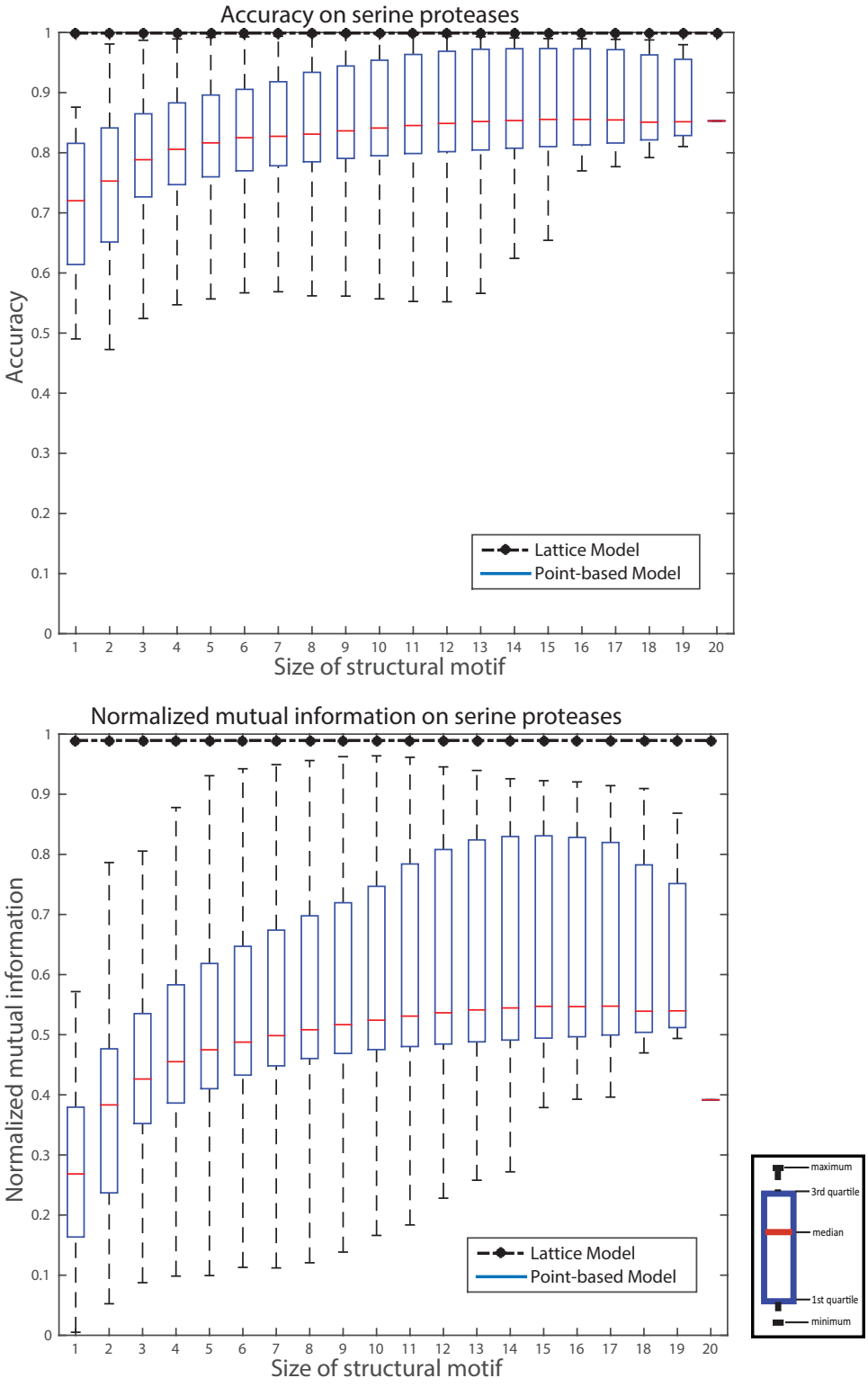

Figure 5.11: Clustering comparison in accuracy (top) and normalized mutual information (bottom) with respect to the number of amino acids in the structural motif on serine proteases.

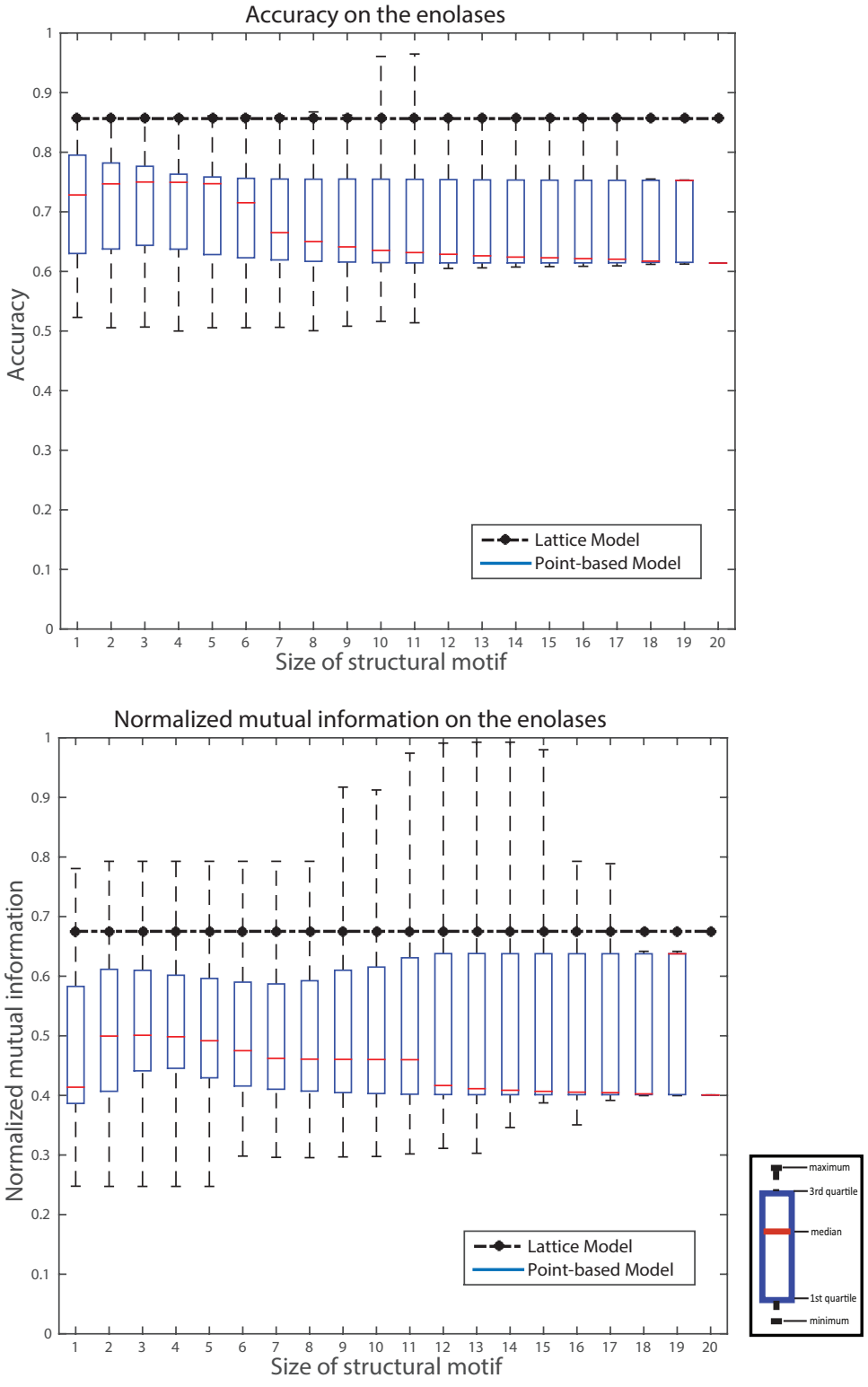

Figure 5.12: Clustering comparison in accuracy (top) and normalized mutual information (bottom) with respect to the number of amino acids in the structural motif on the enolases.

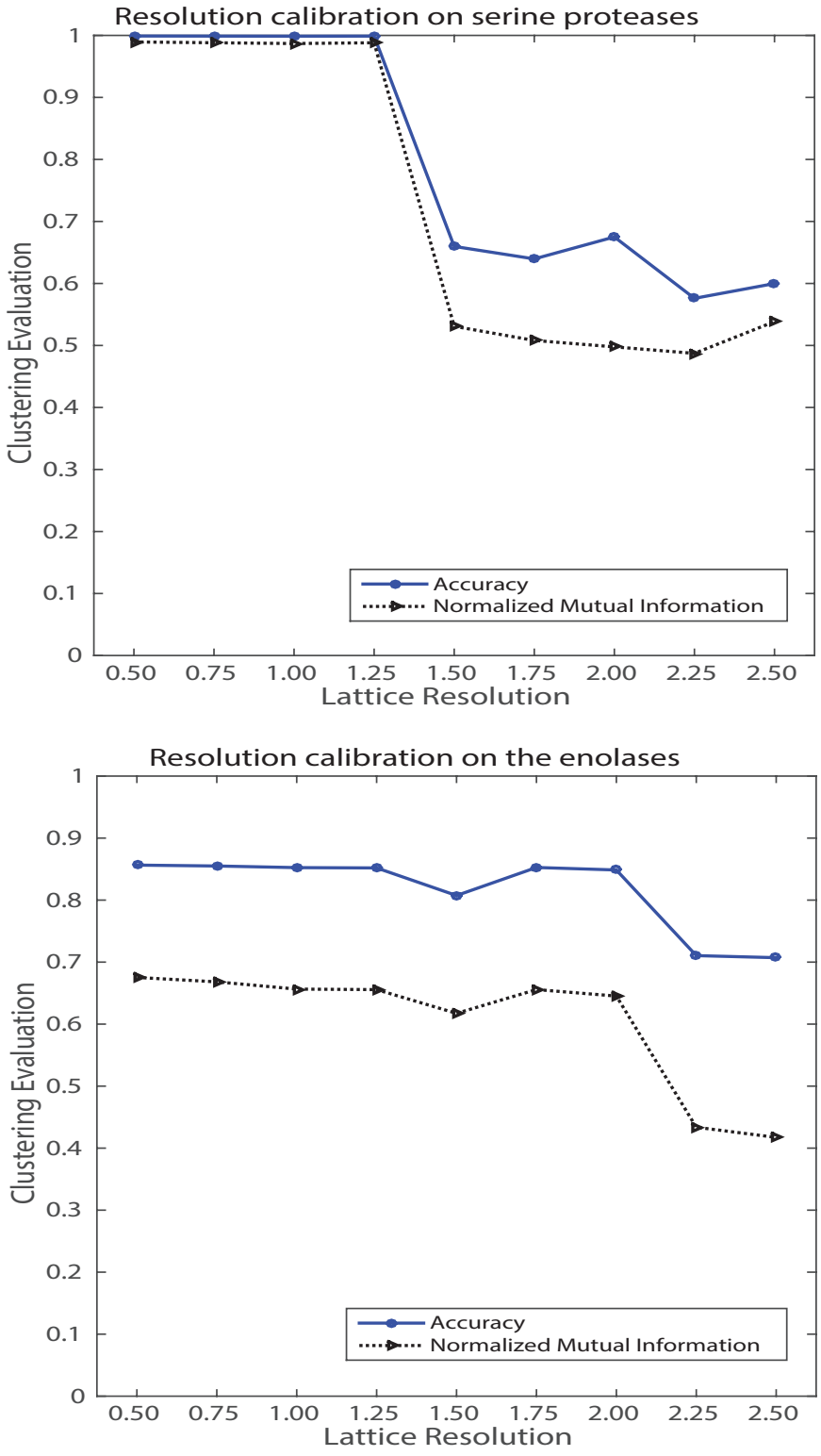

Figure 5.13: The performance of the volumetric lattice representation vs. the lattice resolution r on serine proteases (top) and the enolases (bottom).

First, the volumetric lattice representation outperforms the atomic point representation on both superfamilies. Specifically, on serine proteases, the lattice model performs better than 1048547 (100%) atomic point feature subsets in both accuracy and normalized mutual information evaluation. This shows that, even in the case where the best specificity-sensitive atomic point feature can be found from millions of amino acid combinations, the atomic point representation can only achieve close performance but not as good as the volumetric lattice representation. On the enolases, the volumetric lattice representation performs better than 928641 (88.56%) atomic point feature subsets in clustering accuracy and 903868 (86.20%) atomic point feature subsets in normalized mutual information. All these results prove that volumetric features could learn a better geometric representation of the binding cavity than atomic points. Second, the performance of the atomic point features varies considerably on almost all k values on both superfamilies. This shows that the quality of amino acid subset, which involves in calculating how many amino acids should be selected and which amino acids should be selected, largely influence cavity representation, thus substantially affecting specificity prediction. The volumetric lattice representation, which does not rely on amino acid selection, is obviously more straightforward and user-friendly.

Finally, we analyze the sensitivity to the lattice resolution value. As resolution r decreases, the lattice model approximates the binding cavity using finer cubes, thus leading to more precise representation but computing higher dimensions. Figure 5.13 shows how various resolution values affect specificity prediction. The resolution range between 0.50 and 2.50 is selected in this work because they are more computationally tractable than $r < 0.50$ while they allow for relatively accurate representation. The clustering achieves consistent good performance when the resolution is smaller than 1.50 on serine proteases and when the resolution is smaller than 2.25 on the enolases, respectively.

5.3 An Electrostatic Lattice Representation

Approaches we have introduced so far detect atom coordinates or molecular surface solids for comparative analysis. However, as explained in section 2.3, longer distance electrostatic potentials could selectively affect binding and a comparison of protein electrostatic potentials may explore more depths into specificity analysis. In this section, we will introduce a lattice representation to solve the problem of individual prediction from an electrostatic perspective. It is noted that the electrostatic lattice representation is not intended to analyze dynamics of electrostatic fields because our method does not rely on the order in which the conformations occur.

5.3.1 Method Overview

The electrostatic lattice representation is similar to the volumetric lattice representation, but the core difference is that the electrostatic model builds lattices on protein electrostatic fields rather than molecular surfaces. In this thesis, we calculate electrostatic isopotentials as one way to encode protein electrostatic fields. Electrostatic isopotentials represent geometric surfaces where every point on the surface has the same electrostatic potential, and it was shown that optimizing superposition of electrostatic isopotentials reveals patterns of the binding specificity [154].

First, for each protein conformation, we apply VASP-E [78] to calculate electrostatic isopotentials as solid objects. Second, we describe how we build a lattice model on isopotentials using user-defined cubes. To distinguish from the volumetric lattice representation that computes volumes of binding cavity solids, we called this an electrostatic lattice representation. We then compute the *electrostatic voxel*, which is the isopotential solid volume in each cube, to localize electrostatic charge distributions in the spatial unit. Finally, all electrostatic voxels are combined into a feature vector and the K-means clustering is performed to output specificity prediction on each protein conformation. The overview of the electrostatic lattice representation is shown in Figure 5.14.

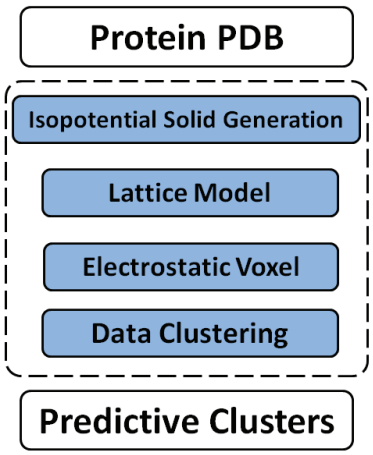

Figure 5.14: The electrostatic lattice representation pipeline.

5.3.2 Solid Representation of Electrostatic Isopotentials

As input, VASP-E requires a protein conformation structure, the electrostatic field and the isopotential threshold $p\ kT/e$. When p is positive, VASP-E represents regions of electrostatic potentials greater than p within a solid region, and when p is negative, regions with potentials less than p are represented. This rule prevents the generation of infinitely large isopotential solids, which lead to degenerate comparisons(Figure 5.15C).

To generate the electrostatic field, we first remove all hydrogens and then protonate the protein structure using the reduce tool of the MolProbity package [155] and the protonated structure is given as input to DelPhi [156] to compute a numerical solution of the Poisson-Boltzmann Equation (PBE). DelPhi approximates the electrostatic field A_E within a bounding box that covers the protein structure. As output, VASP-E generates electrostatic isopotentials as solid objects.

The resulting isopotential solids can have highly convoluted shape. While isopotentials at different thresholds never overlap, they can be formed in close proximity to each other, as can be seen in Figure 5.16C.

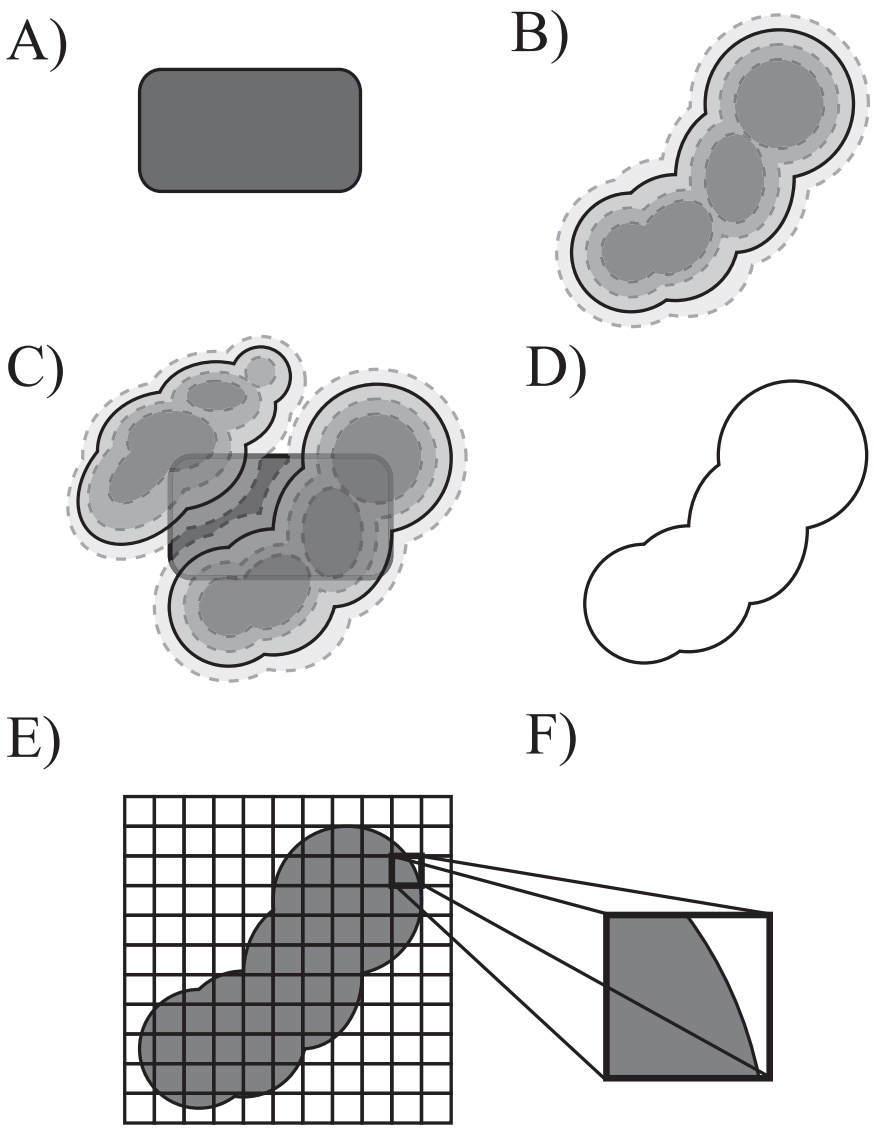

Figure 5.15: An overview of the electrostatic lattice model construction. A) The structure of a given protein conformation. B) The positive electrostatic potentials generated by VASP-E. C) Both positive and negative potentials with respect to the geometric structure. D) The positive electrostatic isopotential selected by k kT/e. E) The bounding box that covers isopotential from all conformations. F) Electrostatic voxel calculation in each lattice cube.

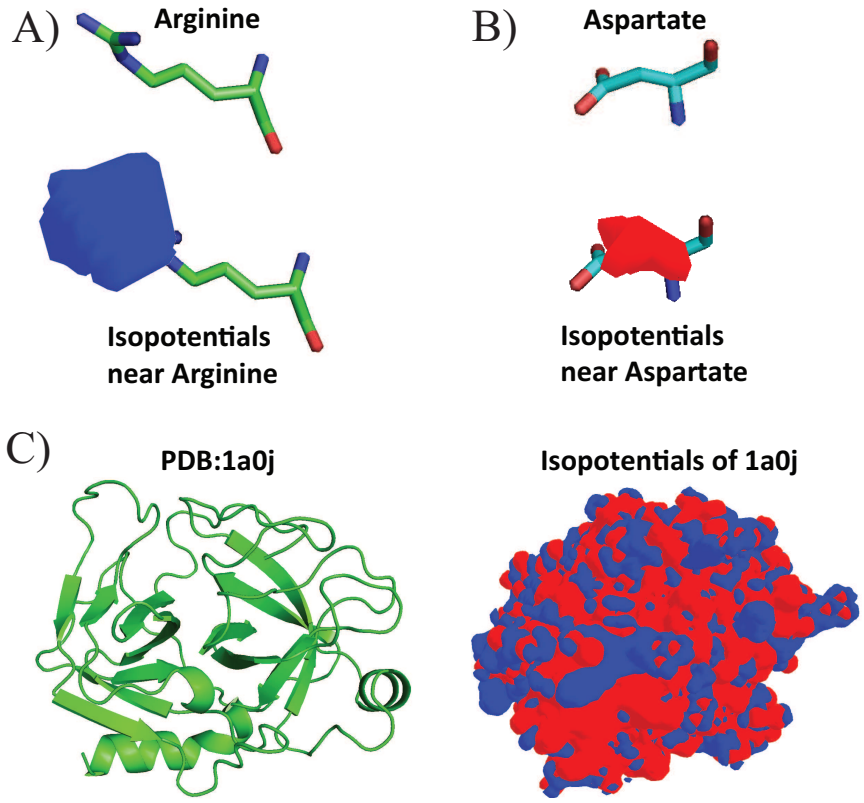

Figure 5.16: A) Electrostatic isopotential of Arginine, a positively charged amino acid, at +2.5 kT/e. B) Electrostatic isopotential of Aspartate, a negatively charged amino acid, at -2.5 kT/e. C) Electrostatic isopotential surfaces of the Atlantic salmon trypsin (pdb:1$a0j$). The red surface indicates the negative isopotential generated at -2.5 kT/e and blue indicates the positive isopotential generated at +2.5 kT/e. The surfaces are highly convoluted and pass very closely to each other, but do not come in contact. The geometric structure of 1$a0j$ is also visualized.

5.3.3 The Lattice Model Construction

The input for constructing the electrostatic lattice are isopotential solids of all protein conformations and a lattice resolution r. Similar to the volumetric lattice representation, a bounding box (Figure 5.15E) is constructed to cover all isopotential solids where the length, the width and the height are all integral multiples of r. Then, we build an axis aligned lattice inside the bounding box so that each cube has the size of r. In the end, we compute the volume of the electrostatic isopotential in each lattice cube (Figure 5.15F).

5.3.4 Cluster Analysis

This step is almost the same as the that of the volumetric lattice representation. The only difference is that we collect all electrostatic voxels rather than volumetric voxels. We also perform the K-means clustering with clustering accuracy (AC) and normalized mutual information evaluation ($\overline{MI}$).

5.3.5 Testing Electrostatic Lattice Representation

We compare the electrostatic lattice representation with geometric structure based methodologies that have been discussed in previous sections. These methods include the catalytic triad or the catalytic pentad, the atomic point representation and the volumetric lattice representation. However, we emphasize here that we are not arguing that the electrostatic based approaches are definitely superior to geometric structure based ones, or vice versa. What we would like to show is that the electrostatic lattice representation is a novel representation and it is comparable to many existing methods. K-means clustering set the number of clusters to be 3 and 20 runs with random initialization were conducted to report the average performance.

Figure 5.17 and Figure 5.18 compare specificity prediction of the electrostatic lattice representation against point-based methods where the isopotential threshold p equals to $+2.5\ kT/e$ because it was shown to be logical selection for isopotential comparisons [154, 78] and the lattice resolution r equals to 2.5. The performance

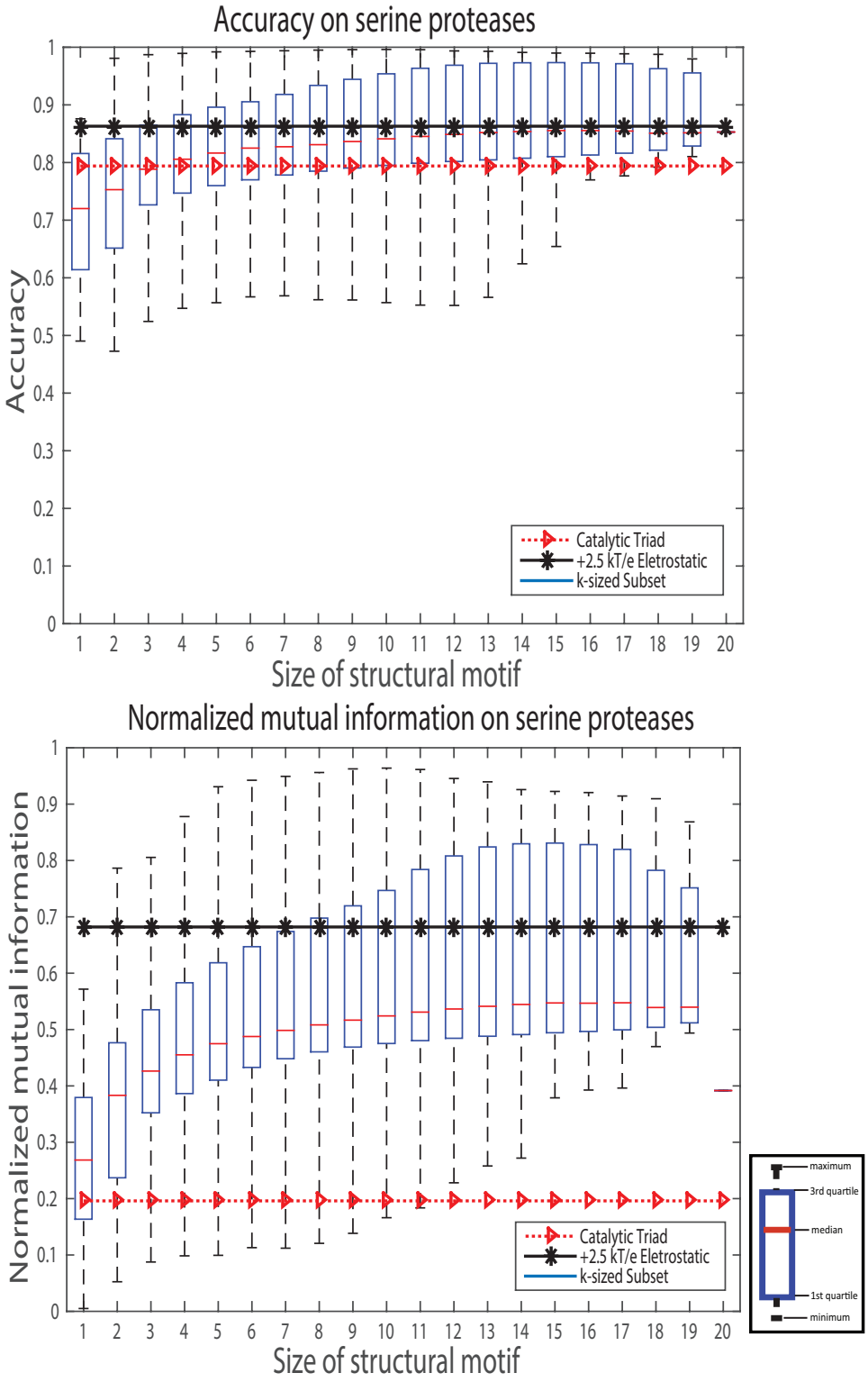

Figure 5.17: Clustering comparison in accuracy (top) and normalized mutual information (bottom) with respect to the number of residues in the structural motif on serine proteases.

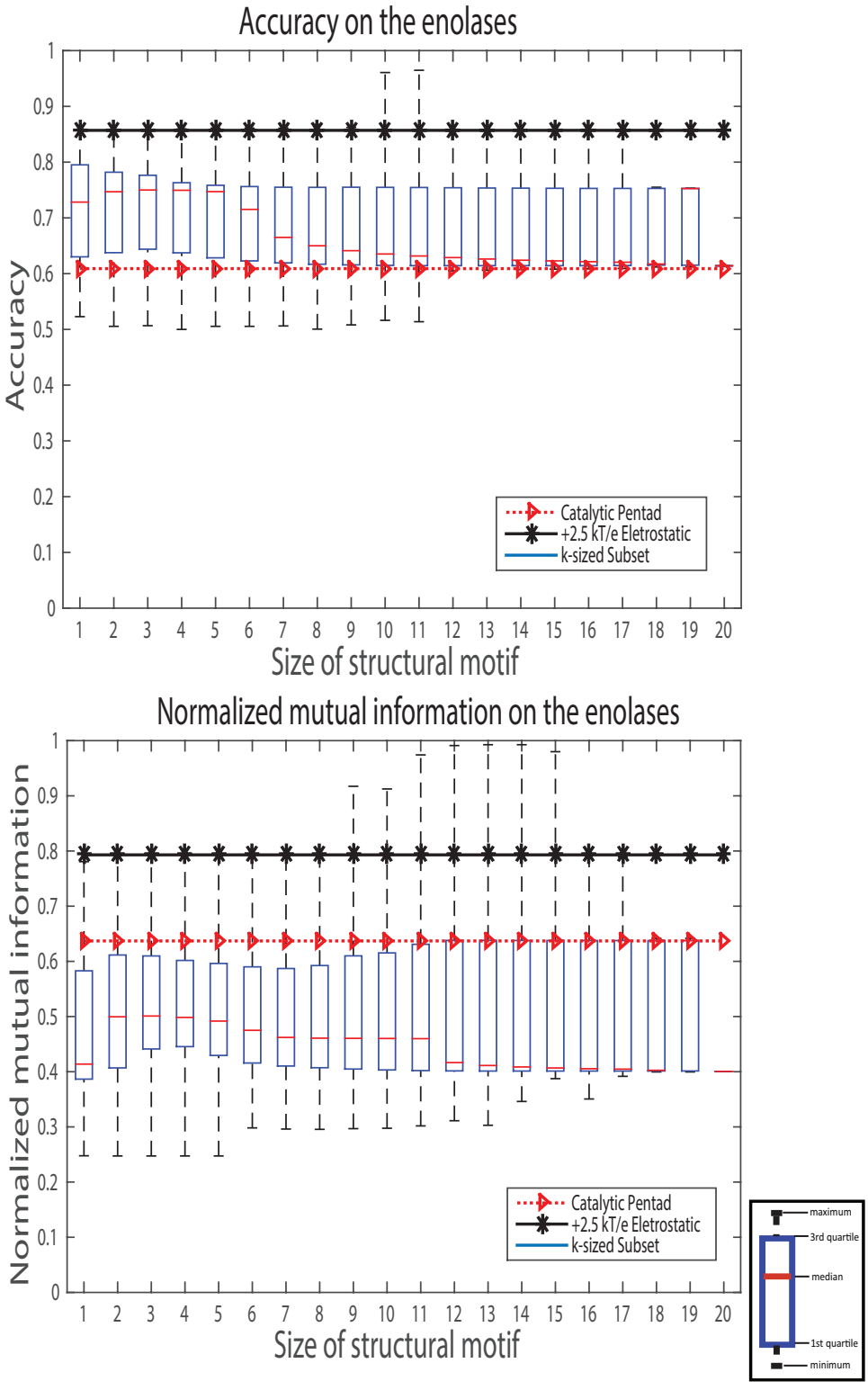

Figure 5.18: Clustering comparison in accuracy (top) and normalized mutual information (bottom) with respect to the number of residues in the structural motif on the enolases.

of k-sized atomic point subset is shown in boxplot on all k-sized feature combinations while the electrostatic lattice representation and the catalytic triad/pentad representation is shown in constant. First, it is obvious that the electrostatic lattice representation outperforms the catalytic triad/pentad on both superfamilies. The catalytic triad and catalytic pentad are well-known examples for characterizing functional sites but they may not be the best selection for predicting specificity. Second, the electrostatic lattice representation outperforms most k-sized atomic point subsets. On serine proteases, it performs better than 586923 (55.97%) atomic point subsets in clustering accuracy and 725974 (69.23%) in normalized mutual information. On the enolases, the electrostatic lattice representation performs better than 965156 (92.05%) atomic point features in clustering accuracy and 1047833 (99.93%) in normalized mutual information. This shows that electrostatic isopotentials could be effective and structure-independent signals for protein comparisons.

We continue to make comparisons to the volumetric lattice representation, electrostatic lattice representations on negative charges and electrostatic representations that concatenate positive with negative features in Table 5.1. The positive electrostatic representation performs as not well as the volumetric representation on serine proteases, but it achieves better performance on the enolases. In addition, The positive electrostatic representation outperforms the negative electrostatic representation and the electrostatic concatenation. One possible explanation could be that the ligand binding of both superfamilies is more affected by positive charges and electrostatic lattice models built on negative charges contain many noises that are not relevant to specificity.

All these results show that the electrostatic lattice representation, which is independent of amino acid selection and structure comparisons, proves to be an effective representation and provides new insights for specificity prediction.

Finally, we conduct sensitivity analysis to the value of the electrostatic lattice resolution. Figure 5.19 shows how the resolution affects clustering performance on $+2.5\ kT/e$ isopotentials where the resolution is ranged from 2.0 to 6.0. The

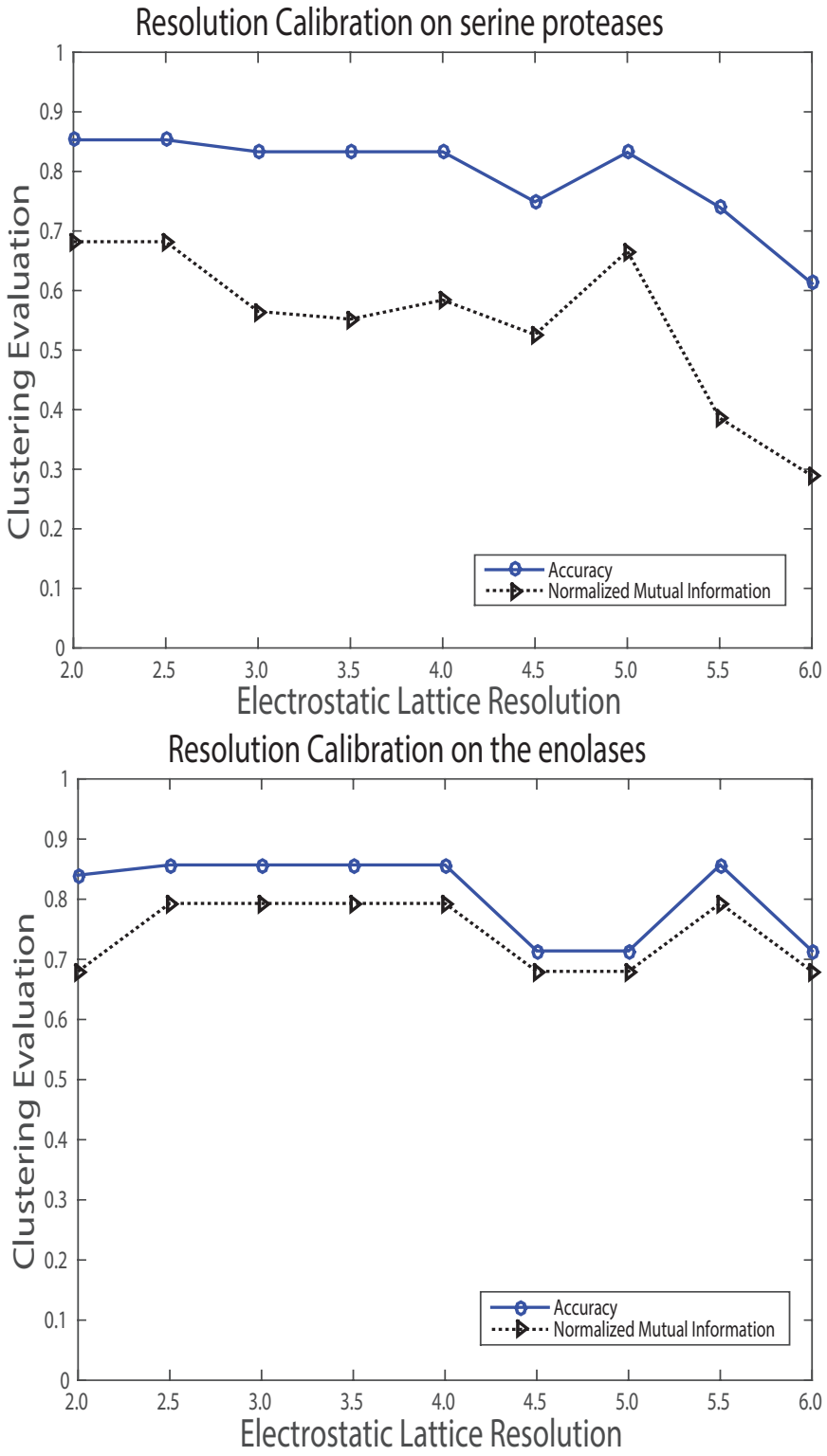

Figure 5.19: The performance of the electrostatic lattice representation vs. the lattice resolution r on serine proteases (top) and the enolases (bottom).

	Serine Proteases			
	$+2.5\ kT/e$	$-2.5\ kT/e$	feature concatenation	volumetric lattice
AC	0.863	0.750	0.749	**0.999**
$\overline{MI}$	0.682	0.422	0.422	**0.989**
	The Enolases			
	$+2.5\ kT/e$	$-2.5\ kT/e$	feature concatenation	volumetric lattice
AC	**0.875**	0.714	0.714	0.856
$\overline{MI}$	**0.793**	0.680	0.682	0.675

Table 5.1: Clustering comparison with volumetric lattice representation and electrostatic lattice representation on negative isopotentials.

resolution value used here is much larger than the value used in the volumetric lattice representation because protein isopotential solids are much larger than binding cavity solids. The performance achieves consistent good performance when the resolution is smaller than 5.0 on serine proteases and on almost all selected resolutions on the enolases.

5.4 Conclusion

In this chapter, we demonstrate three highly modular methods that correspond to the problem of individual prediction. They extract different types of geometric features from either protein structures or protein electrostatic isopotentials to binding specificity on each protein conformation. The atomic point representation identifies coordinates of selected amino acids that are adjacent to the binding cavity. This representation provides the first analysis of map of binding cavity conformations on proteins with different specificities. The volumetric lattice representation extracts volumetric voxels within binding cavity solids, presenting an all-atom motion representation. The electrostatic lattice representation calculates electrostatic voxels to build a lattice model, providing a comparative method that is independent of structure comparisons. By ignoring atomic points or molecular surfaces, the electrostatic lattice representation exclusively reflect conserved or varied regions of electrostatic charge distributions.

The applications of representations investigated in this chapter exist in cases

where individual protein conformations are compared. These representations identify the partner that each conformation will preferentially bind. They are also capable of pointing to structural components, e.i. selected amino acids, user-defined cubes in the binding cavity or the electrostatic potential, that could be altered for the design of a desired specificity.

The substructure matches or lattice construction techniques described in this chapter will facilitate protein comparisons with increasing number of protein structures. The high modularity nature of our methods allow for more customized developments with wider applications beyond protein ligand binding.

Chapter 6

Conclusions and Future Works

This thesis focuses on two predictive problems on binding specificity, aggregate prediction and individual prediction, in the context of protein conformational flexibility. We study two superfamilies, serine proteases and the enolases, of protein structures that exhibit identical folds but different binding specificities.

FAVA demonstrates the first conformationally general tool for predicting specificity by comparing frequent regions of the binding cavity. FAVA also provides the capability to detect amino acids that are influential for specificity. PEAP identifies atomic positions of influential amino acids via motif propagation. The ability of this method to enhance specificity prediction by integrating structural motions in the binding site was demonstrated.

We develop three representative models to solve the the problem of individual prediction. The atomic point representation characterizes the binding cavity with atomic points of selected amino acids. The volumetric lattice representation measures volumes of the binding cavity in user-defined cubes. These two representations reflect geometric changes of binding cavities to compare conformations of different proteins. The electrostatic lattice representation, ignoring atomic points or molecular surface solids, computes volumes of the electrostatic isopotential in user-defined cubes, providing structure-independent techniques for binding analysis.

Together, the methods presented in this thesis leverage protein conformational

flexibility to predict binding specificity. Instead of assuming protein structures to be rigid or partially rigid objects, these methods incorporate diversities of simulated protein conformations, providing more depths into flexible protein comparisons.

The works shown here can be further extended from many directions and we list some that are of particular interest to us. First, the fractional voxels and electrostatic voxels in two lattice models, resembling digital image pixels that compute color values at every physical point, provide an image-like representation for protein comparison. Inspired by the success of convolutional neural networks (CNN) in image recognition [157, 158, 159], we hope that CNN models can also be advantageous to predict binding specificity. CNN is a special type of feed-forward artificial neural network where the individual neuron responds to overlapping regions tiling the local field. A CNN architecture is formed by distinct layers, such as convolutional layers, pooling layers and fully connected layers, and is capable of learning object representation in a increasingly finer way. In the context of specificity prediction, CNN classifiers extract fractional voxels or electrostatic voxels as features to learn CNN parameters from training data sets and output categorical labels to predict binding specificities on testing data sets.

Second, in the electrostatic model, we build lattices on the whole surfaces of electrostatic isopotentials. However, in many cases, selective binding may come from electrostatic effects in a local space, especially near the binding region [106, 113]. To address this issue, we could build a local electrostatic lattice model in the space of binding cavity.

Third, the problem studied in this thesis only output predictive label itself, but ignore partial order between all protein conformations. The partial order is significant because in many application, such as web search, online advertising and recommender systems, we prefer to extract items that are most relevant to input. Similarly, in the context of protein binding, given a ligand as the query, we would like to rank all protein conformations so that conformations of the protein that binds to the ligand should be returned as top results. To deal with this issue, many

learning to rank techniques [160, 161] can be employed.

Bibliography

[1] Pettersen, E. F. *et al.* Ucsf chimeraa visualization system for exploratory research and analysis. *Journal of computational chemistry* **25**, 1605–1612 (2004).

[2] DeLano, W. L. The pymol molecular graphics system (2002). URL `https://www.pymol.org/`.

[3] Garrett, R. & Grisham, C. M. Biochemistry, 1999. *Saunders College Publishing* .

[4] Parker, D. C. T cell-dependent b cell activation. *Annual review of immunology* **11**, 331–360 (1993).

[5] Prochiantz, A. Messenger proteins: homeoproteins, tat and others. *Current opinion in cell biology* **12**, 400–406 (2000).

[6] Wheeler, D. L. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **35**, D5–D12 (2007).

[7] Bork, P. *et al.* Predicting function: from genes to genomes and back. *Journal of molecular biology* **283**, 707–725 (1998).

[8] Webb, E. C. *et al. Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Ed. 6 (Academic Press, 1992).

[9] Keseler, I. M. *et al.* Ecocyc: a comprehensive database resource for escherichia coli. *Nucleic acids research* **33**, D334–D337 (2005).

[10] Moszer, I., Glaser, P. & Danchin, A. Subtilist: a relational database for the bacillus subtilis genome. *Microbiology* **141**, 261–268 (1995).

[11] Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29 (2000).

[12] Marcotte, E. M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).

[13] Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research* **28**, 33–36 (2000).

[14] Wu, J.-S., Huang, S.-J. & Zhou, Z.-H. Genome-wide protein function prediction through multi-instance multi-label learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **11**, 891–902 (2014).

[15] Jones, P. *et al.* Interproscan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

[16] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences* **96**, 4285–4288 (1999).

[17] Qian, B. & Goldstein, R. A. Detecting distant homologs using phylogenetic tree-based hmms. *Proteins: Structure, Function, and Bioinformatics* **52**, 446–453 (2003).

[18] Engelhardt, B. E., Jordan, M. I. & Brenner, S. E. A graphical model for predicting protein molecular function. In *Proceedings of the 23rd international conference on Machine learning*, 297–304 (ACM, 2006).

[19] Jiang, D., Pei, J., Ramanathan, M., Tang, C. & Zhang, A. Mining coherent gene clusters from gene-sample-time microarray data. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 430–439 (ACM, 2004).

[20] Ng, S.-K., Zhu, Z. & Ong, Y.-S. Whole-genome functional classification of genes by latent semantic analysis on microarray data. In *Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29*, 123–129 (Australian Computer Society, Inc., 2004).

[21] Huynen, M. A., Snel, B., von Mering, C. & Bork, P. Function prediction and protein networks. *Current opinion in cell biology* **15**, 191–198 (2003).

[22] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. & Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21**, i302–i310 (2005).

[23] Couto, F. M., Silva, M. J. & Coutinho, P. Profal: Protein functional annotation through literature. In *JISBD*, 747–756 (2003).

[24] Koike, A., Niwa, Y. & Takagi, T. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics* **21**, 1227–1236 (2005).

[25] Altschul, S. F. *et al.* Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389–3402 (1997).

[26] Finn, R. D., Clements, J. & Eddy, S. R. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research* gkr367 (2011).

[27] Servant, F. *et al.* Prodom: automated clustering of homologous domains. *Briefings in bioinformatics* **3**, 246–251 (2002).

[28] Dobson, P. D., Cai, Y.-D., Stapley, B. J. & Doig, A. J. Prediction of protein function in the absence of significant sequence similarity. *Current medicinal chemistry* **11**, 2135–2142 (2004).

[29] Liu, X. Deep recurrent neural network for protein function prediction from sequence. *arXiv preprint arXiv:1701.08318* (2017).

[30] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. Genbank. *Nucleic acids research* **33**, D34–D38 (2005).

[31] Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *Journal of molecular biology* **307**, 1113–1143 (2001).

[32] Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **5**, 823 (1986).

[33] Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequencea study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics* **77**, 499–508 (2009).

[34] Huang, I. K., Pei, J. & Grishin, N. V. Defining and predicting structurally conserved regions in protein superfamilies. *Bioinformatics* **29**, 175–181 (2013).

[35] Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).

[36] Holm, L. & Sander, C. Mapping the protein universe. *Science* **273**, 595–602 (1996).

[37] Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein engineering* **11**, 739–747 (1998).

[38] Zhang, Y. & Skolnick, J. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research* **33**, 2302–2309 (2005).

[39] Wallace, A. C., Borkakoti, N. & Thornton, J. M. Tess: a geometric hashing algorithm for deriving 3d coordinate templates for searching structural

databases. application to enzyme active sites. *Protein science* **6**, 2308–2323 (1997).

[40] Wang, C. & Scott, S. D. New kernels for protein structural motif discovery and function classification. In *Proceedings of the 22nd international conference on Machine learning*, 940–947 (ACM, 2005).

[41] Bryant, D. H., Moll, M., Finn, P. W. & Kavraki, L. E. Combinatorial clustering of residue position subsets predicts inhibitor affinity across the human kinome. *PLoS computational biology* **9**, e1003087 (2013).

[42] Dundas, J. *et al.* Castp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic acids research* **34**, W116–W118 (2006).

[43] Chen, B. Y. & Honig, B. VASP: A volumetric analysis of surface properties yields insights into protein-ligand binding specificity. *PLoS computational biology* **6**, e1000881 (2010).

[44] Chen, B. Y. & Bandyopadhyay, S. Vasp-s: A volumetric analysis and statistical model for predicting steric influences on protein-ligand binding specificity. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, 22–29 (IEEE, 2011).

[45] Anfinsen, C. Principles that govern the protein folding chains. *Science* **181**, 233–230 (1973).

[46] Ramachandran, G. K. *et al.* A bond-fluctuation mechanism for stochastic switching in wired molecules. *Science* **300**, 1413–1416 (2003).

[47] Bu, Z., Cook, J. & Callaway, D. J. Dynamic regimes and correlated structural dynamics in native and denatured alpha-lactalbumin. *Journal of molecular biology* **312**, 865–873 (2001).

[48] Bu, Z. & Callaway, D. Proteins move! protein dynamics and long-range allostery in cell signaling. *Adv Protein Chem Struct Biol* **83**, 163–221 (2011).

[49] Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).

[50] Chou, K.-C. & Shen, H.-B. Hum-ploc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochemical and biophysical research communications* **347**, 150–157 (2006).

[51] Yu, G., Domeniconi, C., Rangwala, H., Zhang, G. & Yu, Z. Transductive multi-label ensemble classification for protein function prediction. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1077–1085 (ACM, 2012).

[52] Freitas, A. A., Wieser, D. C. & Apweiler, R. On the importance of comprehensible classification models for protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **7**, 172–182 (2010).

[53] Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology* **233**, 123–138 (1993).

[54] Nussinov, R. & Wolfson, H. J. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proceedings of the National Academy of Sciences* **88**, 10495–10499 (1991).

[55] Orengo, C. A. & Taylor, W. R. Ssap: sequential structure alignment program for protein structure comparison. *Computer methods for macromolecular sequence analysis* (1996).

[56] Petrey, D. & Honig, B. Grasp2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods in enzymology* **374**, 492–509 (2003).

[57] Yang, A.-S. & Honig, B. An integrated approach to the analysis and modeling of protein sequences and structures. i. protein structural alignment and a quantitative measure for protein structural distance. *Journal of molecular biology* **301**, 665–678 (2000).

[58] Xie, L. & Bourne, P. E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. *Proceedings of the National Academy of sciences* **105**, 5441–5446 (2008).

[59] Gibrat, J.-F., Madej, T. & Bryant, S. H. Surprising similarities in structure comparison. *Current opinion in structural biology* **6**, 377–385 (1996).

[60] Chen, B. Y. *et al.* The mash pipeline for protein function prediction and an algorithm for the geometric refinement of 3d motifs. *Journal of Computational Biology* **14**, 791–816 (2007).

[61] Barker, J. A. & Thornton, J. M. An algorithm for constraint-based structural template matching: application to 3d templates with statistical analysis. *Bioinformatics* **19**, 1644–1649 (2003).

[62] Russell, R. B. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *Journal of molecular biology* **279**, 1211–1227 (1998).

[63] Chen, B. Algorithms for structural comparison and statistical analysis of 3d protein motifs by chen, vy fofanov, dm kristensen, m. kimmel, o. lichtarge, and le kavraki pacific symposium on biocomputing 10: 334-345 (2005). In *Pacific Symposium on Biocomputing*, vol. 10, 334–345 (Citeseer, 2005).

[64] Shatsky, M., Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. Recognition of binding patterns common to a set of protein structures. In *Research in Computational Molecular Biology*, 440–455 (Springer, 2005).

[65] Chen, B. Y. *et al.* Cavity-aware motifs reduce false positives in protein function prediction. In *Proceedings of the 2006 IEEE Computational Systems Bioinformatics Conference (CSB 2006)*, 311–23 (2006).

[66] Porter, C. T., Bartlett, G. J. & Thornton, J. M. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic acids research* **32**, D129–D133 (2004).

[67] Hulo, N. *et al.* The prosite database. *Nucleic acids research* **34**, D227–D230 (2006).

[68] Attwood, T. K. *et al.* Prints and its automatic supplement, preprints. *Nucleic acids research* **31**, 400–402 (2003).

[69] Chen, B. Y. *et al.* Cavity scaling: automated refinement of cavity-aware motifs in protein function prediction. *Journal of bioinformatics and computational biology* **5**, 353–382 (2007).

[70] Bryant, D. H., Moll, M., Chen, B. Y., Fofanov, V. Y. & Kavraki, L. E. Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction. *BMC bioinformatics* **11**, 242 (2010).

[71] Kristensen, D. M. *et al.* Prediction of enzyme function based on 3d templates of evolutionarily important amino acids. *BMC bioinformatics* **9**, 17 (2008).

[72] Chen, B. Y. *et al.* Geometric sieving: Automated distributed optimization of 3d motifs for protein function prediction. In *Annual International Conference on Research in Computational Molecular Biology*, 500–515 (Springer, 2006).

[73] Guo, Z. & Chen, B. Y. Variational bayesian clustering on protein cavity conformations for detecting influential amino acids. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 703–710 (ACM, 2014).

[74] Honig, B. & Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **268**, 1144 (1995).

[75] Nakamura, H. Roles of electrostatic interaction in proteins. *Quarterly reviews of biophysics* **29**, 1–90 (1996).

[76] Rosen, M., Lin, S. L., Wolfson, H. & Nussinov, R. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Engineering* **11**, 263–277 (1998).

[77] Kinoshita, K. & Nakamura, H. Identification of the ligand binding sites on the molecular surface of proteins. *Protein Science* **14**, 711–718 (2005).

[78] Chen, B. Y. Vasp-e: Specificity annotation with a volumetric analysis of electrostatic isopotentials. *PLoS Comput Biol* **10** (2014).

[79] Binkowski, T. A., Adamian, L. & Liang, J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *Journal of molecular biology* **332**, 505–526 (2003).

[80] Laskowski, R. A. Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics* **13**, 323–330 (1995).

[81] Ritchie, D. W. & Kemp, G. J. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *Journal of Computational Chemistry* **20**, 383–395 (1999).

[82] Kazhdan, M., Funkhouser, T. & Rusinkiewicz, S. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, vol. 6, 156–164 (2003).

[83] Dobson, P. D. & Doig, A. J. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology* **330**, 771–783 (2003).

[84] Nayal, M. & Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins: Structure, Function, and Bioinformatics* **63**, 892–906 (2006).

[85] Dobson, P. D. & Doig, A. J. Predicting enzyme class from protein structure without alignments. *Journal of molecular biology* **345**, 187–199 (2005).

[86] Syed, U. & Yona, G. Using a mixture of probabilistic decision trees for direct prediction of protein function. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, 289–300 (ACM, 2003).

[87] Borgwardt, K. M. *et al.* Protein function prediction via graph kernels. *Bioinformatics* **21**, i47–i56 (2005).

[88] Bhattacharya, S., Bhattacharyya, C. & Chandra, N. Structural alignment based kernels for protein structure classification. In *Proceedings of the 24th international conference on Machine learning*, 73–80 (ACM, 2007).

[89] Qiu, J., Hue, M., Ben-Hur, A., Vert, J.-P. & Noble, W. S. A structural alignment kernel for protein structures. *Bioinformatics* **23**, 1090–1098 (2007).

[90] Novinskaya, A., Devaurs, D., Moll, M. & Kavraki, L. E. Improving protein conformational sampling by using guiding projections. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, 1272–1279 (IEEE, 2015).

[91] Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nature methods* **10**, 221–227 (2013).

[92] Ye, Y. & Godzik, A. Fatcat: a web server for flexible structure comparison and structure similarity searching. *Nucleic acids research* **32**, W582–W585 (2004).

[93] Gunasekaran, K. & Nussinov, R. How different are structurally flexible and rigid binding sites? sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *Journal of molecular biology* **365**, 257–273 (2007).

[94] Shatsky, M., Nussinov, R. & Wolfson, H. J. Flexprot: alignment of flexible protein structures without a predefinition of hinge regions. *Journal of Computational Biology* **11**, 83–106 (2004).

[95] Ye, Y. & Godzik, A. Multiple flexible structure alignment using partial order graphs. *Bioinformatics* **21**, 2362–2369 (2005).

[96] Konc, J. & Janežič, D. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **26**, 1160–1168 (2010).

[97] Mosca, R. & Schneider, T. R. Rapido: a web server for the alignment of protein structures in the presence of conformational changes. *Nucleic acids research* **36**, W42–W46 (2008).

[98] Menke, M., Berger, B. & Cowen, L. Matt: local flexibility aids protein multiple structure alignment. *PLoS computational biology* **4**, e10 (2008).

[99] Vesterstrøm, J. & Taylor, W. R. Flexible secondary structure based protein structure comparison applied to the detection of circular permutation. *Journal of Computational Biology* **13**, 43–63 (2006).

[100] Kmiecik, S. *et al.* Coarse-grained protein models and their applications. *Chemical Reviews* **116**, 7898–7936 (2016).

[101] Senn, H. M. & Thiel, W. Qm/mm methods for biomolecular systems. *Angewandte Chemie International Edition* **48**, 1198–1229 (2009).

[102] Binder, K. Monte carlo simulations in statistical physics. In *Encyclopedia of Complexity and Systems Science*, 5667–5677 (Springer, 2009).

[103] Sharp, K. A. & Honig, B. Calculating total electrostatic energies with the nonlinear poisson-boltzmann equation. *Journal of Physical Chemistry* **94**, 7684–7692 (1990).

[104] Gilson, M. K., Sharp, K. A. & Honig, B. H. Calculating the electrostatic potential of molecules in solution: method and error assessment. *Journal of computational chemistry* **9**, 327–335 (1988).

[105] Livesay, D. R., Jambeck, P., Rojnuckarin, A. & Subramaniam, S. Conservation of electrostatic properties within enzyme families and superfamilies. *Biochemistry* **42**, 3464–3473 (2003).

[106] Murray, D. & Honig, B. Electrostatic control of the membrane targeting of c2 domains. *Molecular cell* **9**, 145–154 (2002).

[107] Gilson, M. K. & Honig, B. H. Calculation of electrostatic potentials in an enzyme active site. *Nature* **330**, 84–86 (1987).

[108] McCoy, A. J., Epa, V. C. & Colman, P. M. Electrostatic complementarity at protein/protein interfaces. *Journal of molecular biology* **268**, 570–584 (1997).

[109] Botti, S. A., Felder, C. E., Sussman, J. L. & Silman, I. Electrotactins: a class of adhesion proteins with conserved electrostatic and structural motifs. *Protein engineering* **11**, 415–420 (1998).

[110] Richard, A. M. Quantitative comparison of molecular electrostatic potentials for structure-activity studies. *Journal of computational chemistry* **12**, 959–969 (1991).

[111] Zhang, X. *et al.* Application of new multiresolution methods for the comparison of biomolecular electrostatic properties in the absence of global structural similarity. *Multiscale Modeling & Simulation* **5**, 1196–1213 (2006).

[112] Kinoshita, K. & Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. *Protein Science* **12**, 1589–1595 (2003).

[113] Kinoshita, K., Murakami, Y. & Nakamura, H. ef-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic acids research* **35**, W398–W402 (2007).

[114] Morihara, K. & Tsuzuki, H. Comparison of the specificities of various serine proteinases from microorganisms. *Archives of biochemistry and biophysics* **129**, 620–634 (1969).

[115] Graf, L. *et al.* Electrostatic complementarity within the substrate-binding pocket of trypsin. *Proceedings of the National Academy of Sciences* **85**, 4961–4965 (1988).

[116] Berglund, G. I., Smalas, A. O., Outzen, H. & Willassen, N. P. Purification and characterization of pancreatic elastase from north atlantic salmon (salmo salar). *Molecular marine biology and biotechnology* **7**, 105–114 (1998).

[117] Babbitt, P. C. *et al.* The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the α-protons of carboxylic acids. *Biochemistry* **35**, 16489–16501 (1996).

[118] Kühnel, K. & Luisi, B. F. Crystal structure of the escherichia coli rna degradosome component enolase. *Journal of molecular biology* **313**, 583–592 (2001).

[119] Schafer, S. L. *et al.* Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the d270n mutant. *Biochemistry* **35**, 5662–5669 (1996).

[120] Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation* **4**, 435–447 (2008).

[121] Berendsen, H., Postma, J., van Gunsteren, W. & Hermans, J. Intermolecular forces. *Pullman, B., Ed.; Reidel Publishing Company: Dordrecht* 331–342 (1981).

[122] Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics* **52**, 7182 (1981).

[123] Nose, S. & Klein, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics* **50**, 1055–1076 (1983).

[124] Hess, B. P-lincs: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation* **4**, 116–122 (2008).

[125] Voelcker, H. B. & Requicha, A. A. Geometric modeling of mechanical parts and processes. *Computer* **10**, 48–57 (1977).

[126] Petrey, D. & Honig, B. Grasp2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods in enzymology* **374**, 492–509 (2002).

[127] Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–713 (1983).

[128] Guo, Z. *et al.* A flexible volumetric comparison of protein cavities can reveal patterns in ligand binding specificity. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 445–454 (ACM, 2014).

[129] Biggiogero, G. La geometria del tetraedro. *Enciclopedia delle Matematiche Elementari e Complementi* **2**, 219–252 (1950).

[130] Sneath, P. H. & Sokal, R. R. *Numerical taxonomy. The principles and practice of numerical classification.* (1973).

[131] Junier, T. & Zdobnov, E. M. The newick utilities: high-throughput phylogenetic tree processing in the unix shell. *Bioinformatics* **26**, 1669–1670 (2010).

[132] Schechter, I. & Berger, A. On the size of the active site in proteases. I. Papain. *Biochemical and Biophysical Research Communications* **27**, 157–162 (1967).

[133] Shotton, D. & Watson, H. Three-dimensional structure of tosyl-elastase. *Nature* **225**, 811–816 (1970).

[134] Rokach, L. Ensemble-based classifiers. *Artificial Intelligence Review* **33**, 1–39 (2010).

[135] Zhou, Z.-H. *Ensemble methods: foundations and algorithms* (CRC Press, 2012).

[136] Laskowski, R. A., Watson, J. D. & Thornton, J. M. Protein function prediction using local 3d templates. *Journal of molecular biology* **351**, 614–626 (2005).

[137] Strehl, A. & Ghosh, J. Cluster ensembles-a knowledge reuse framework for combining partitionings. In *AAAI/IAAI*, 93–99 (2002).

[138] Moll, M., Bryant, D. H. & Kavraki, L. E. The labelhash algorithm for substructure matching. *BMC bioinformatics* **11**, 1 (2010).

[139] Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

[140] Lee, D. D. & Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 556–562 (2001).

[141] Jolliffe, I. *Principal component analysis* (Wiley Online Library, 2002).

[142] Lovász, L. & Plummer, M. D. *Matching theory*, vol. 367 (American Mathematical Soc., 2009).

[143] Henschel, A., Kim, W. K. & Schroeder, M. Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics* **22**, 550–555 (2006).

[144] Meng, E. C., Polacco, B. J. & Babbitt, P. C. Superfamily active site templates. *PROTEINS: Structure, Function, and Bioinformatics* **55**, 962–976 (2004).

[145] Kleywegt, G. J. Recognition of spatial motifs in protein structures. *Journal of molecular biology* **285**, 1887–1897 (1999).

[146] Pegg, S. *et al.* Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry* **45**, 2545–2555 (2006).

[147] Orengo, C. A. *et al.* Cath–a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1109 (1997).

[148] Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* **247**, 536–540 (1995).

[149] Xu, W., Liu, X. & Gong, Y. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 267–273 (ACM, 2003).

[150] Shahnaz, F., Berry, M. W., Pauca, V. P. & Plemmons, R. J. Document clustering using nonnegative matrix factorization. *Information Processing & Management* **42**, 373–386 (2006).

[151] Ding, C., Li, T., Peng, W. & Park, H. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 126–135 (ACM, 2006).

[152] Cai, D., He, X., Han, J. & Huang, T. S. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**, 1548–1560 (2011).

[153] Schaer, J. & Stone, M. Face traverses and a volume algorithm for polyhedra. In *New Results and New Trends in Computer Science*, 290–297 (Springer, 1991).

[154] Guo, Z., Scheinberg, K., Hong, J. & Chen, B. Y. Superposition of protein structures using electrostatic isopotentials. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, 75–82 (IEEE, 2015).

[155] Chen, V. B. *et al.* Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66**, 12–21 (2009).

[156] Rocchia, W., Alexov, E. & Honig, B. Extending the applicability of the nonlinear poisson-boltzmann equation: Multiple dielectric constants and multivalent ions. *The Journal of Physical Chemistry B* **105**, 6507–6514 (2001).

[157] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).

[158] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105 (2012).

[159] Ji, S., Xu, W., Yang, M. & Yu, K. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **35**, 221–231 (2013).

[160] Liu, T.-Y. *et al.* Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* **3**, 225–331 (2009).

[161] Li, H. Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies* **7**, 1–121 (2014).

Biography

Ziyi Guo was born on August 5, 1990 in Jiangsu, China. He studied at Xuzhou No.1 Middle School from 2002 to 2008. He then studied and obtained a bachelor degree in Software Engineering from Northwestern Polytechnical University in 2012 where he worked with Prof. Lei Xie on natural language processing on the final year project. After that, he enrolled in Computer Science and Enigeering Department, Lehigh University, for a Ph.D. study in Computer Science. At Lehigh, he worked with Prof. Brian Chen at Informatics Lab. His research interests are structural bioinformatics and applied machine learning.